LECTURES ON PROBABILITY THEORY
Math GU 4155 Spring 2021

IOANNIS KARATZAS *

October 4, 2022

Abstract

We develop the basic notions and results of Probability Theory including probability spaces, events, conditional expectations, independence, the Laws of Large Numbers, the Central Limit Theorems, Random Walk, and the theory of MARKOV Chains.

* Department of Mathematics, Columbia University, New York, NY 10027. E-Mail: *ik1@columbia.edu*. I am indebted to my colleagues Professors Duong PHONG, Peter BANK, Robert NEEL, Julien DUBÉDAT, Ivan CORWIN and Evgeni DIMITROV, for making available to me notes for classes they had taught over the years at Columbia. Their notes helped me greatly in putting this collection of topics together.

Contents

1 Introduction

Probability Theory is the branch of Mathematics concerned with the analysis of random phenomena: those whose outcomes cannot be predicted with certainty in advance. Its name has Latin roots: *probare* (to prove, or test) and *ilis* (to be able). [1] [2]

Probability is considerably younger than most other areas of Mathematics, having originated as a mathematical discipline with the correspondence between Blaise PASCAL and Pierre DE FERMAT in 1654 on the study of games of chance; see the popular book by DEVLIN (2008) for historical perspective. For a wonderful exposition of the basic mathematical, philosophical and scientific aspects of the subject, as well as its rich history, consult DIACONIS & SKYRMS (2018). [3]

Very influential in the development of the subject in that early stage, was the appearance of the first monograph on Probability *"De Ratiociniis in Ludo Aleae"* by Christiaan HUYGENS (1657), as well as the posthumous publication of the *"Ars Conjectandi"* by Jacob BERNOULLI (1713). [4]

Since then, in the three centuries that have followed, Probability Theory has become indispensable for dealing with randomness in almost all branches of science. It is now essential in the discovery and study of macroscopic regularities that occur when large systems of particles, organ-

[1] Randomness was known to the ancients. Egyptians and Babylonians played games of chance, as did Homeric heroes (see Figure 3) and the Romans. The Greeks had *Tyche,* the Goddess of chance, and used the same word for "happiness" and for "good luck"; the Romans instituted a lottery to help fill state coffers; soldiers cast a lot for the cloak of the Christ. Democritus and his followers, Epicurus in particular, postulated a physical source of randomness, the "swerve" in the *"De Rerum Natura"* of Lucretius, which affects all the atoms that make up the Universe. This Lucretian "swerve" (*clinamen* in Latin) is what we call "Brownian Motion" today.

But the ancients, both Greek and Roman, were also rather uncomfortable with randomness (just as much, perhaps, as they feared, and tried to avoid, infinity). They tended to view the physical world as subject to immutable, unambiguous laws, and randomness just as the manifestation of man's inability to come to grips with the true nature of events. It never occurred to the Romans, for instance, that it might be possible to quantify the revenue they could expect to raise from their lottery, and to plan the finances of the Eternal City and of its Empire accordingly. Over the centuries, Christian scholars struggled and wrestled with the conflict between the belief in free will and its implied randomness, and the idea that God knows everything that happens.

[2] *Probabilism* also refers to the philosophical doctrine of Academic Skepticism which goes back to the "skeptical" period of ancient Platonism that start in the third century BC. The main thesis here, is that in the absence of certainty, plausibility (or verisimilitude, or the assessment of the odds) is the best possible path to knowledge.

[3] One of the most influential polymaths of the Renaissance, Hieronymus CARDANUS (Gerolamo CARDANO, 1501-1576), was apparently the first to formulate the idea that chance can in fact be *measured,* and to enunciate the notion of what we call today "independence". His *Liber de Ludo Aleae* is, according to DEVLIN (2008), "the first scientific study of dice rolling, based on the premise that there are fundamental principles governing the likelihood of particular outcomes". The book is partly observational (he had a lot of opportunity for observation, long hours at the gambling table); and partly a theoretical analysis of how chance events, such as particular outcomes of rolls of dice, aggregate when repeated many times. In modern parlance, it was the first study of frequentist probability." Perennially short of money, CARDANO would keep himself solvent by gambling and chess-playing. He made many contributions to mathematics, physics, biology, chemistry, and astronomy – including the invention/perfection of the compass and of the gyroscope. Years after CARDANO's death, in the early 17th century the great Galileo GALILEI (1564-1642) made very similar observations trying to answer questions about games of chance posed to him by his patron, the Great Duke of Tuscany.

[4] The two-volume work by BELL (1937) contains excellent biographical and mathematical sketches for many of the main actors in the early development of our subject, including PASCAL, FERMAT, HUYGENS, BERNOULLI, LAPLACE, GAUSS and CANTOR, among many others. For the mathematical lineage of KOLMOGOROV and the Moscow mathematical school founded by his teachers SUSLIN and LUZIN, and its ties to the French school of BOREL, LEBESGUE, BAIRE and to the field of descriptive set theory, see the fascinating tract by GRAHAM & KANTOR (2009).

isms, or agents, interact according to the laws of physics, chemistry, biology, or economics; in the study of population genetics and of the genome; in the study of polymers, and of interface growth or random deposition phenomena; in the study of ferromagnetism or percolation near criticality; in the study of how signals are transmitted through a noisy channel, and then recovered; in the theory and practice of finance and insurance; in the design and analysis of large-scale communication, neural or queueing networks, and of algorithms for combinatorial optimization, computerized tomography, signal processing, pattern recognition; and so on.

The mathematical foundations of the subject, as we know and study it now, were laid out in the monograph by A.N. KOLMOGOROV (1933). Today, in addition to its position as a full-fledged area within Mathematics – and its interactions with Analysis, Partial Differential Equations, Geometry, Combinatorics, Algebra, Topology, Number Theory, Representation Theory and Mathematical Physics – Probability forms the language of Statistics and of the Quantitative Social Sciences. [5]

A central rôle in several of these applications is played by the development of Stochastic Analysis over the last 70 years starting with the pioneering work of ITÔ (1942, 1944) and DOOB (1953). Probability has also been central in the study of Finance, ever since BACHELIER (1900) pioneered the mathematical study of Brownian Motion and understood its significance as a tool for the analysis of financial markets – five years before EINSTEIN (1905) developed his physical theory for the phenomenon known as "Brownian movement", which was then studied from the mathematical point of view by WIENER (1923, 1924) and LÉVY (1948).

Additional developments abound. Let us just mention the entire field of Information Theory, which started with the seminal book by SHANNON & WEAVER (1949); the explosion in the study of interacting particle systems of Statistical Physics, starting with the pioneering work of SPITZER (1969) and LIGGETT (1985); and the tremendous advances in the study of random matrices, inspired by statistical theory, data analysis, and models for heavy-nuclei atoms (WISHART (1928), WIGNER (1958), DYSON (1962), DYSON & MEHTA (1963), ANDERSON ET AL. (2010)).

1.1 Preview

Our aim in this course will be to provide a mathematically rigorous introduction to the basic notions and results of Probability Theory (chapter 2). We shall introduce probability spaces and random variables, including several examples that we shall revisit throughout the course (chapter 3), and we shall study relevant aspects of measure theory (chapters 4 and 5). We shall show how to construct probability spaces with desirable properties and characteristics (chapter 6), and how to deal with the all-important notions of conditional probabilities and expectations (chapter 8). We shall develop basic concepts such as independence, as well as fundamental results like the BOREL-CANTELLI lemmata and the KOLMOGOROV 0-1 law (chapter 7). We shall study in detail the

[5] By contrast, the physical, algorithmic and philosophical underpinnings of the subject, including the question of what constitutes "true randomness", are still very much studied and debated vigorously; see the book by SHAFER & VOVK (2019) for a modern account of these issues, including the "frequentist" and "collective" approaches of VON MISES (1919) and VILLE (1939), and the newer, game-theoretic ones. It is very interesting that KOLMOGOROV himself pioneered in the 1950's an approach to randomness based on the concept of "maximal algorithmic complexity" that he also introduced; see KOLMOGOROV (1983) and MARTIN-LÖF (1966) for accounts of this approach. For an up-to-date account of the role of randomness in algorithmic complexity and computation, consult WIGDERSON (2019).

simple, symmetric random walk (chapter 9). Next, we shall study various modes of convergence for random variables, and prove the law of large numbers and the central limit theorem (chapters 10, 11). Finally, we shall introduce the theory of MARKOV chains (chapter 9).

An effort will be made to bring forth both the rich intuitive content of the subject *and* its mathematical underpinnings and subtleties. For this reason we shall rely on two textbooks: the book by STIRZAKER (2003) is a rich source of examples and applications, whereas the book by WALSH (2012) will serve as the theoretical backbone of the course. No student should miss the chance to delve, at some point, into one of the most important treatises written on this subject – the two volumes by W. FELLER (1968, 1971).

A recurrent theme throughout the course will be that relying on intuition alone, without the discipline and rigor of the subject's mathematical foundations, can lead very easily to grave errors; we shall try to illustrate a few of these in the form of "paradoxes", sprinkled throughout the text.

2 Notions

For the study of random phenomena we need models which, invariably, require the construction of what we call *Probability Space.*

There are three things required to define such a space.

• First, we need the set Ω of all possible outcomes of the experiment we have in mind; we call this set *Sample Space*. For instance:

If we throw two dice, the sample space is $\Omega_1 = \{(i,j) : 1 \leq i,j \leq 6\}$ with $6 \times 6 = 36$ elements.

If we keep track of the number of cars passing by a certain point on the highway during a given time interval, say one hour, then $\Omega_2 = \mathbb{N}_0$ is the set of nonnegative integers.

If we keep tossing a coin *ad infinitum*, the sample space is the collection of strings of 0's (tails) and 1's (heads), namely $\Omega_3 = \big\{(\omega_1, \omega_2, \cdots) : \omega_i \in \{0,1\}\big\} = \{0,1\}^{\mathbb{N}}$.

If we throw a dart on the unit interval $(0,1)$, then the set of all possible real numbers that we can hit is $\Omega_4 = (0,1)$.

• Secondly, we need to decide which subsets of the sample space are important to us – to the extent that, eventually, we would like to "measure" them by assigning "weights" or probabilities to them, and to have a calculus for computing these probabilities. We denote by $\mathcal{F}$ the collection of all such "measurable" subsets, which we call *Events.* [6]

For instance, in the case of a sample space Ω of finite cardinality like Ω_1 above, we can consider taking as "events" the collection

$$\mathcal{F} = \mathcal{P}(\Omega)$$

of *all* subsets of the sample space: its "power set". We can consider doing the same also in the case of Ω of infinite but countable cardinality, like $\Omega_2 = \mathbb{N}_0$.

In the case of $\Omega_3 = \{0,1\}^{\mathbb{N}}$, we could consider taking its power set, but this may contain more sets than we care about. In this case we definitely want our $\mathcal{F}$ to contain the collection $\mathcal{C}$ of "cylinder sets" $\{\omega : (\omega_1, \cdots, \omega_k) = (\varepsilon_1, \cdots, \varepsilon_k)\}$ for arbitrary but fixed integer k and $\varepsilon_i \in \{0,1\},\ i = 1, \cdots, k$.

In the case of $\Omega_4 = (0,1)$, we definitely want $\mathcal{F}$ to contain all intervals of the form (a,b) for $0 \leq a < b \leq 1$.

• Finally, we need a recipe for assigning probabilities to all the events under consideration; in other words, we need to specify a mapping

$$\mathcal{F} \ni A \longmapsto \mathbb{P}(A) \in [0,1]$$

that assigns to each event a number in [0,1], its *Probability*.

Quite clearly and intuitively, probabilities cannot be negative or larger than 100%.

[6] *"Events, young man, events"*, was the answer of the Right Honourable Harold MACMILLAN, Earl of Stockton and British Prime Minister in the late 1950's and early 1960's, to a journalist's question of what made his work so unpredictable and hard.

For instance, if Ω has finite cardinality, we may wish to assign to each subset A of the sample space the probability

$$\mathbb{P}(A) = \frac{|A|}{|\Omega|}\,, \tag{2.1}$$

to wit, the ratio of its cardinality divided by the cardinality of the sample space. This probability measure is called the "normalized counting measure", and corresponds to the intuitive notion that "all elements of the sample space are *equally likely*":

$$\mathfrak{p}_\omega = \mathbb{P}\big(\{\omega\}\big) = \frac{1}{|\Omega|}$$

for every $\omega \in \Omega$. For instance, $\mathfrak{p}_\omega = 1/36$ in the case of Ω_1 above.

A bit more generally, if the cardinality of Ω is finite, or even countably infinite, we can take $\mathcal{F} = \mathcal{P}(\Omega)$, fix a collection of nonnegative numbers $\mathfrak{p}_\omega$ with $\sum_{\omega\in\Omega} \mathfrak{p}_\omega = 1$, and define

$$\mathbb{P}(A) \;=\; \sum_{\omega\in A} \mathfrak{p}_\omega$$

for arbitrary $A \in \mathcal{P}(\Omega)$. For instance, in the case of $\Omega_2 = \mathbb{N}_0$ above, we can take the so-called POISSON distribution with parameter $\lambda > 0$, given by

$$\mathfrak{p}_\omega \;=\; e^{-\lambda}\,\frac{\lambda^\omega}{\omega!}\,, \quad \omega = 0, 1, 2, \cdots\,.$$

Rules: Needless perhaps to say, the choice of both $\mathcal{F}$ (the collection of events) and $\mathbb{P} : \mathcal{F} \to [0,1]$ (the rule for assigning probabilities to these events) has to be subject to certain rules.

For starters, we should like $\emptyset \in \mathcal{F}$, $\Omega \in \mathcal{F}$ and $\mathbb{P}(\emptyset) = 0$, $\mathbb{P}(\Omega) = 1$. We would also like the union $A \cup B$ of two *disjoint* ($A \cap B = \emptyset$) events in $\mathcal{F}$ to have probability equal to the sum

$$\mathbb{P}(A \cup B) \;=\; \mathbb{P}(A) + \mathbb{P}(B)$$

of the probabilities of the individual events – much like with the notions of "area" or "volume". Once we have it for *two* events, this *additivity property* holds also for any *finite* number of pairwise-disjoint events:

$$\mathbb{P}\Big(\bigcup_{k=1}^{n} A_k\Big) \;=\; \sum_{k=1}^{n} \mathbb{P}(A_k)$$

for every integer $n \in \mathbb{N}$ and every collection of events $\{A_k\}_{k=1}^n \subseteq \mathcal{F}$ that are *pairwise disjoint,* that is, satisfy $A_i \cap A_j = \emptyset$ for $1 \le i \ne j \le n$.

It turns out that, for very important technical reasons, we should like to have this additivity property also for *countable* unions of pairwise disjoint subsets of the sample space Ω. [7]

[7] It is certainly possible to construct theories of measure, probability and integration that are based on finite, as opposed to countable, additivity. Foremost among these is the magisterial work of DUBINS & SAVAGE (1965), who developed such a theory of probability in the context of stochastic optimization. The resulting notions of integral and expectation, however, are not easy to work with: they do not interact well with limiting operations, that is, they lack the very properties (monotone convergence, dominated convergence, FATOU's lemma, etc.) that make the LEBESGUE integral such a useful tool for both theory and applications.

But then, if Ω is not countable, say as in $\Omega_4 = (0,1)$, this simply *cannot be done* in general, for an arbitrary collection $\mathcal{F}$ of subsets of Ω. It is just not possible to construct a countably additive set function $\mathbb{P} : \mathcal{P}((0,1)) \to [0,1]$ that assigns measure $\mathbb{P}((c,d)) = d - c$ to every subinterval $(c,d) \subseteq (0,1)$, within any system for Mathematics that is compatible with the Axiom of Choice; see Exercise 4.14 below.

More generally, it is impossible to construct a nonnegative set function, defined on *all* the subsets of the unit cube in Euclidean space, which is countably additive, assigns full measure to the unit cube, and assigns the same measure to sets that are "congruent" (i.e., can be transformed into each other by means of translation, rotation and/or reflection). This is the so-called BANACH & TARSKI (1924) "paradox".

Even more to the point, perhaps, it is often *not even desirable* to take $\mathcal{F}$ as the collection of all subsets of the sample space. Consider an experiment, such as the evolution of the price of a certain asset, that occurs progressively in time. As time t evolves in $\mathbb{N}$ or $\mathbb{R}_+$, we need to base investment decisions on a sub-collection $\mathcal{F}_t$ of events in $\mathcal{F}$, which correspond to the information available to us up to and including that time – and not beyond, since typically we should expect $\mathcal{F}_t \subset \mathcal{F}_u$ for $t < u$. Similar considerations apply also when it comes to the concept of "conditional probability" in chapter 8.

Thus, we shall require that a family $\mathcal{F}$ of subsets of Ω must be specified, and must have some specific properties consistent with the definition of measure, namely those of a *σ-algebra*. The properties of σ-algebras and of measures are self-evident enough if we consider only finite unions of measurable sets. It is their strengthening to *countable* unions which accounts for the power and flexibility of the theory; but also for the difficulties encountered in the construction of measures.

We formalize these considerations in the next section.

2.1 Sigma-Algebras and Measures

Consider a nonempty set Ω, our Sample Space. We have the following two very important notions.

Definition 2.1. *A collection $\mathcal{F}$ of subsets of Ω is called* algebra (*respectively,* σ-algebra)*, if it contains Ω and is closed under complementation and under finite (resp., countable) unions.*

If $\mathcal{F}$ is a σ-algebra *of subsets of Ω, the pair $(\Omega, \mathcal{F})$ is called* Measurable Space*.*

It is clear that a $\sigma-$algebra contains the empty set, and is closed under countable intersections as well: from DE MORGAN's laws, we have

$$\left(\bigcap_{k \in \mathbb{N}} A_k \right)^c = \bigcup_{k \in \mathbb{N}} A_k^c \,, \qquad \left(\bigcup_{k \in \mathbb{N}} A_k \right)^c = \bigcap_{k \in \mathbb{N}} A_k^c \,.$$

Concrete examples appear in the next section.

We shall construct measures, including probability measures, on σ-algebras of subsets of the sample space. Thus, the domains on which measures are defined are closed under set-operations such as complementation, countable unions, countable intersections, countable differences.

The notion of measure respects these operations in a very nice way. I recall a talk by Gian-Carlo ROTA in the early 1980's at Columbia, during which he called this "the triumph of definition".

Definition 2.2. Measure and Probability: *Given a σ-algebra $\mathcal{F}$ of subsets of Ω as in Definition 2.1, a set function $\mathbb{P} : \mathcal{F} \to [0, \infty]$ will be called* measure *if it satisfies $\mathbb{P}(\emptyset) = 0$ and is countably additive, i.e.,*

$$\mathbb{P}\Big(\bigcup_{k\in\mathbb{N}} A_k\Big) \;=\; \sum_{k\in\mathbb{N}} \mathbb{P}(A_k)$$

holds for any countable collection $\{A_k\}_{k\in\mathbb{N}}$ of pairwise disjoint sets in $\mathcal{F}$.

We shall call a measure finite, *if $\mathbb{P}(\Omega) < \infty$; we shall say that a measure space as above is σ-*finite, *if there is a countable partition $\Omega = \bigcup_{j\in\mathbb{N}} A_j$ of the sample space by sets $\{A_j\}_{j\in\mathbb{N}} \subset \mathcal{F}$ with $\mathbb{P}(A_j) < \infty$ for all $j \in \mathbb{N}$.*

A finite measure will be called probability measure, *if $\mathbb{P}(\Omega) = 1$; then the triple $(\Omega, \mathcal{F}, \mathbb{P})$ as in Definition 2.1 and above, is called* Probability Space.

Several consequences of this definition are immediate:

- For arbitrary events $A_1, A_2, \cdots$ on a measure space $(\Omega, \mathcal{F}, \mathbb{P})$ we have the *Countable Subadditivity Property*

$$\mathbb{P}\Big(\bigcup_{k\in\mathbb{N}} A_k\Big) \;\le\; \sum_{k\in\mathbb{N}} \mathbb{P}(A_k)\,.$$

- Monotonicity: $\mathbb{P}(A) \le \mathbb{P}(B)\,, \quad$ if $\; A \subseteq B \;$ and $\;\; A \in \mathcal{F}\,,\; B \in \mathcal{F}\,.$
- Continuity from below: If $A_1 \subseteq A_2 \subseteq \cdots$, then

$$\lim_{k\to\infty} \uparrow \mathbb{P}(A_k) = \mathbb{P}\Big(\bigcup_{k\in\mathbb{N}} A_k\Big)\,.$$

- Continuity from above: If $A_1 \supseteq A_2 \supseteq \cdots$ and $\mathbb{P}(A_1) < \infty$, then

$$\lim_{k\to\infty} \downarrow \mathbb{P}(A_k) = \mathbb{P}\Big(\bigcap_{k\in\mathbb{N}} A_k\Big)\,.$$

- $\mathbb{P}(A \cup B) + \mathbb{P}(A \cap B) = \mathbb{P}(A) + \mathbb{P}(B)\,, \quad \forall \;\; A \in \mathcal{F}\,,\; B \in \mathcal{F}\,.$
- On a probability space we have: $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)\,, \quad \forall \;\; A \in \mathcal{F}\,.$
- For arbitrary events $A_1, \cdots, A_n$ on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ we have the so-called *Inclusion-Exclusion Formula*

$$\boxed{\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) \;=\; \sum_{i=1}^{n} \mathbb{P}(A_i) - \sum_{i<j} \mathbb{P}(A_i \cap A_j) + \sum_{i<j<k} \mathbb{P}(A_i \cap A_j \cap A_k) \;-\; \cdots \;+\; (-1)^{n+1}\, \mathbb{P}\left(\bigcap_{i=1}^{n} A_i\right)}$$

as well as the BONFERRONI *Inequalities*

$$\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) \;\le\; \sum_{i=1}^{n} \mathbb{P}(A_i)\,, \quad \mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) \;\ge\; \sum_{i=1}^{n} \mathbb{P}(A_i) - \sum_{i<j} \mathbb{P}(A_i \cap A_j)\,,$$

$$\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) \;\le\; \sum_{i=1}^{n} \mathbb{P}(A_i) - \sum_{i<j} \mathbb{P}(A_i \cap A_j) + \sum_{i<j<k} \mathbb{P}(A_i \cap A_j \cap A_k)$$

and so on: if we stop the inclusion/exclusion formula of the right-hand side after an even (respectively, odd) number of steps, we get a lower (respectively, upper) bound.

Remark 2.1. Given any nonempty set Ω, we have the *trivial σ-algebra* $\mathcal{F} = \{\emptyset, \Omega\}$ which is pitifully small, as well as the "large" σ-algebra $\mathcal{F} = \mathcal{P}(\Omega)$ that consists of all subsets of Ω. This last σ-algebra is typically very big, when Ω has infinitely (in particular, uncountably) many elements.

Exercise 2.1. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and consider a set $\mathcal{J}$ which is (at most) countable. For every $j \in \mathcal{J}$ we are given two events $B_j \subseteq A_j$ in $\mathcal{F}$. Show that

$$\mathbb{P}\Big(\bigcup_{j\in\mathcal{J}} A_j\Big) - \mathbb{P}\Big(\bigcup_{j\in\mathcal{J}} B_j\Big) \leq \sum_{j\in\mathcal{J}} \big[\,\mathbb{P}(A_j) - \mathbb{P}(B_j)\,\big]\,.$$

2.2 Smallest Sigma Algebra

Let us make a very simple, yet also very important, observation: The intersection $\bigcap_{\alpha\in I} \mathcal{F}_\alpha$ of an arbitrary collection $\{\mathcal{F}_\alpha\}_{\alpha\in I}$ of σ-algebras is a σ-algebra.

Now, for any given family $\mathcal{G}$ of subsets of Ω, let us consider the collection of all σ-algebras that contain it; this collection is nonempty, as it contains the σ-algebra $\mathcal{P}(\Omega)$. Thus, the intersection of all the σ-algebras of this collection is a σ-algebra, indeed the smallest σ-algebra that contains $\mathcal{G}$. We shall denote it by $\sigma(\mathcal{G})$, and call it the *σ-algebra generated by the family* $\mathcal{G}$. [8]

For instance, in the case of Ω_3 above, we want to take $\mathcal{F} = \sigma(\mathcal{C})$, the smallest sigma algebra generated by the collection $\mathcal{C}$ of cylinder sets.

Whereas, in the the case of Ω_4 above, we definitely want to take $\mathcal{F}$ to be the smallest sigma algebra generated by the collection of all intervals of the form (a, b) for $0 \leq a < b \leq 1$. This leads to the very important notion of BOREL sets.

2.3 BOREL Sets

Definition 2.3. BOREL Sets; Measurable Functions: *Let us consider a metric space $\mathbb{S}$, as well as the collection $\mathcal{O}$ of its open sets.*

(i) The collection $\mathcal{B}(\mathbb{S})$ of BOREL sets of $\mathbb{S}$ *is the smallest σ-algebra that contains the open sets:*

$$\mathcal{B}(\mathbb{S}) := \sigma(\mathcal{O})\,.$$

*(ii) A function $X : \Omega \to \mathbb{S}$ is then called $\mathcal{F}-$*measurable [9] *if for every* BOREL *set $B \in \mathcal{B}(\mathbb{S})$ we have*

$$X^{-1}(B) := \big\{\omega \in \Omega \,:\, X(\omega) \in B\big\} \in \mathcal{F}\,;$$

or more compactly, $X^{-1}(\mathcal{B}(\mathbb{S})) \subseteq \mathcal{F}$.

Likewise, a real-valued function $F : \mathbb{S} \to \mathbb{R}$ is called BOREL measurable, *if $F^{-1}(A) \in \mathcal{B}(\mathbb{S})$ holds for every* BOREL *set $A \in \mathcal{B}(\mathbb{R})$; or more compactly, $F^{-1}(\mathcal{B}(\mathbb{R})) \subseteq \mathcal{B}(\mathbb{S})$.*

[8] It is useful to keep in mind that $\sigma(\mathcal{G})$ contains all the sets you can obtain, if you start with countably many sets from $\mathcal{G}$, and employ countably many set-theoretic operations of unions, intersections or complementations.

[9] Or simply "measurable", when there is no scope for confusion regarding the $\sigma-$algebra of subsets, with which the sample space is endowed.

It is not hard to see from this definition that every continuous, real-valued function $F : \mathfrak{S} \to \mathbb{R}$ defined on a metric space $\mathfrak{S}$, is BOREL measurable. [10] The reverse is clearly not true: for instance, every increasing, right continuous function $F : \mathbb{R} \to \mathbb{R}$ is BOREL measurable, but may have plenty of (in fact, countably many) discontinuity points.

Remark 2.2. *Two Important Properties of Inverse Functions:* For *any* function $X : \Omega \to \mathfrak{S}$ and any index set $\mathcal{I}$ (not necessarily countable), we have $X^{-1}(A^c) = \big(X^{-1}(A)\big)^c$ and

$$X^{-1}\Big(\bigcup_{\alpha\in\mathcal{I}} A_\alpha\Big) = \bigcup_{\alpha\in\mathcal{I}} X^{-1}(A_\alpha)\,, \qquad X^{-1}\Big(\bigcap_{\alpha\in\mathcal{I}} A_\alpha\Big) = \bigcap_{\alpha\in\mathcal{I}} X^{-1}(A_\alpha)\,. \tag{2.2}$$

In particular: the pre-image under X of a σ-algebra of subsets of $\mathfrak{S}$, is a σ-algebra of subsets of Ω.

Remark 2.3. *The* BOREL *Sets of the Real Line:* In the case of the real line, the BOREL σ-algebra $\mathcal{B}(\mathbb{R})$ is also generated by half-lines or by intervals, for instance of the type $(-\infty, a]$, (a, ∞) or $(a, b]$, with $-\infty < a < b < \infty$.

As a consequence, a function $F : \mathfrak{S} \to \mathbb{R}$ is BOREL measurable, iff $F^{-1}\left((-\infty, a]\right) \in \mathcal{B}(\mathfrak{S})$ holds for every $a \in \mathbb{R}$; equivalently, iff $F^{-1}\left((a, \infty)\right) \in \mathcal{B}(\mathfrak{S})$ holds for every $a \in \mathbb{R}$.

There exist subsets of the real line that are not BOREL. Such sets are necessarily uncountable. We shall "meet" some of them in due course – but once again rely on the Axiom of Choice for the intruductions.

Remark 2.4. *Algebra, But Not $\sigma-$Algebra:* The collection $\mathcal{G}$ of "left-closed/right-open" intervals $\big\{[a, b) : 0 \le a \le b \le 1\big\}$ is *not* a σ-algebra of subsets of $\Omega = [0, 1)$: it easy to check that $\bigcup_{n\in\mathbb{N}} [1/n, 1) = (0, 1) \notin \mathcal{G}$.

It is also straightforward to see that the collection $\mathcal{E}$ of all finite disjoint unions of such intervals is an *algebra*. Of course, there is a smallest σ-algebra $\sigma(\mathcal{G})$ that contains this collection – namely, the BOREL σ-algebra, and $\sigma(\mathcal{E}) = \sigma(\mathcal{G})$.

Exercise 2.2. Sigma Algebras Generated by Functions: Let X, Y be real-valued, $\mathcal{F}-$measurable functions on the space $(\Omega, \mathcal{F})$.

(i) Observe that the collection

$$\sigma(X) := X^{-1}\big(\mathcal{B}(\mathbb{R})\big)$$

is a σ-algebra, and satisfies $\sigma(X) \subseteq \mathcal{F}$. It is called the *$\sigma$-algebra generated by X*, and is the smallest σ-algebra of subsets of Ω, with respect to which X is measurable.

(ii) (J.L. DOOB) For $\sigma(Y) \subseteq \sigma(X)$ to hold, it is necessary and sufficient that there exist a BOREL-measurable function $h : \mathbb{R} \to \mathbb{R}$ such that

$$Y(\omega) = h\big(X(\omega)\big)\,, \quad \omega \in \Omega\,.$$

[10] The collection $\mathcal{C} := \{B \subseteq \mathbb{R} : F^{-1}(B) \in \mathcal{B}(\mathfrak{S})\}$ is easily seen to be a σ-algebra; and it contains all open subsets of $\mathbb{R}$ (because, by the assumed continuity of F, the pre-image of every open subset of $\mathbb{R}$ is an open set in $\mathfrak{S}$, that is, an element of $\mathcal{O}$). Thus $\mathcal{C}$ also contains $\mathcal{B}(\mathbb{R})$, the smallest σ-algebra containing the collection $\mathcal{O}$ of open sets.

(iii) A bit more generally, given a collection $\mathcal{C}$ of real-valued, measurable functions on $(\Omega, \mathcal{F})$, we shall denote by $\sigma(\mathcal{C})$, and call it *the $\sigma-$algebra generated by the collection $\mathcal{C}$*, the smallest $\sigma-$algebra of sets in $\mathcal{F}$ with respect to which every element of this collection is measurable.

Exercise 2.3. Let Ω be an arbitrary nonempty set, and denote by $\mathcal{C}$ the collection of all its "singletons", that is, all subsets that consist of exactly one element of Ω. Show that

$$\sigma(\mathcal{C}) = \mathcal{A} := \{ A \subset \Omega : A \text{ or } A^c \text{ is countable} \}.$$

2.4 Some Examples of Measure and Probability Spaces

Here are some examples of probability spaces. We shall construct more elaborate examples in the sections and chapters to follow.

(a): The space for n *coin tosses;* this space $\Omega = \{0,1\}^n$ has $N = 2^n$ elements, conveniently represented as $n-$tuples $\omega = (\omega_1, \cdots, \omega_n)$ with $\omega_i \in \{0,1\}$. Each ω_i can be interpreted as the outcome of the i-th coin toss, with $\omega_i = 1$ when the outcome is "heads" and $\omega_i = 0$ when the outcome is "tails". This space is endowed with the collection $\mathcal{F}$ of all its subsets and, say, with normalized counting measure as in (2.1).

(b): More generally, given any set Ω of finite cardinality $|\Omega| < \infty$, we take $\mathcal{F} = \mathcal{P}(\Omega)$ to be the collection of all $2^{|\Omega|}$ subsets; we also specify nonnegegative numbers $\mathfrak{p}_\omega$ with $\sum_{\omega \in \Omega} \mathfrak{p}_\omega = 1$. Then the recipe below defines a probability measure on $(\Omega, \mathcal{F})$:

$$\mathbb{P}(A) := \sum_{\omega \in A} \mathfrak{p}_\omega \qquad \text{for any } \ A \subset \Omega .$$

When we take $\mathfrak{p}_\omega = 1/|\Omega|$, that is, we consider all elements of Ω "equally likely", this recipe gives us normalized counting measure $\mathbb{P}(A) = |A|/|\Omega|$.

(c): The DIRAC *measure* δ_c at the point $c \in \mathbb{R}^d$, on the Euclidean space $\Omega = \mathbb{R}^d$ with all subsets of Ω measurable, and with $\delta_c(E)$ equal to either 1 or 0, depending on whether the set E contains the point c or not.

To illustrate the subsequent discussion, we shall also assume for the moment that

(d): there exists a unique measure λ on the $\sigma-$algebra $\mathcal{F} = \mathcal{B}(\mathbb{R})$ of BOREL subsets of the real line, which coincides with the notion of length when restricted to intervals:

$$\lambda(I) = b - a, \ \text{ when } \ I = (a, b].$$

This measure λ is called LEBESGUE *measure* on $\mathbb{R}$. If this construction is restricted to the unit interval [0,1], then the resulting measure $\lambda\,|_{[0,1]}$ is a probability measure.

We have the following generalization. Let us agree to call *distribution function* every $F : \mathbb{R} \to \mathbb{R}$ which is nondecreasing and right-continuous; if in addition we have $F(-\infty) = 0$ and $F(\infty) = 1$, we call this F a *probability distribution function*.

(e): For every distribution function F there exists a unique measure μ_F defined on a σ-algebra containing all intervals in $\mathbb{R}$, so that

$$\mu_F(I) = F(b) - F(a) \ \text{ when } \ I = (a, b].$$

This measure is called the LEBESGUE-STIELTJES *measure* associated with F, and it becomes LEBESGUE measure when $F(x) = x$; whereas, if F is a probability distribution function, this μ_F is a probability measure.

We shall construct such measures in the following chapters. [11] The construction depends on a celebrated result known as **Karathéodory Extension Theorem,** which can be stated for our current purposes very roughly as follows: *"Any $\sigma-$finite measure on an algebra $\mathcal{G}$ can be extended, and uniquely, to a measure on the $\sigma-$algebra $\mathcal{F} = \sigma(\mathcal{G})$."*

2.5 Measurable Functions, Again

Given a measurable space $(\Omega, \mathcal{F})$, a function $X : \Omega \to \mathbb{R}$ is measurable (with respect to $\mathcal{F}$) if all pre-images of open half-lines

$$X^{-1}((a, \infty)) \equiv \{\omega \in \Omega \,:\, X(\omega) > a\}$$

belong to the σ-algebra $\mathcal{F}$, for every $a \in \mathbb{R}$; recall Remark 2.3.

Taking complements, countable unions and intersections, we deduce that $X^{-1}(I) \in \mathcal{F}$ holds for any interval I. Useful here are (2.2), and the observations

$$[a, \infty) \,=\, \bigcap_{n\in\mathbb{N}} (a - 1/n\,, \infty) \;,$$

$$[a, b] = \bigcap_{n\in\mathbb{N}} (a - 1/n\,, b + 1/n) \;, \qquad (a, b) = \bigcup_{n\in\mathbb{N}} [a + 1/n\,, b - 1/n] \;;$$

these show that any $\sigma-$algebra that contains the open intervals also contains the closed ones, and vice-versa.

Exercise 2.4. Preservation of measurability under simple operations. Let X, Y be real-valued, measurable functions of $(\Omega, \mathcal{F})$. If c is a real number, show that the functions below are also measurable:

$$c\,X\,, \qquad X^2\,, \qquad X + Y\,, \qquad X\,Y\,, \qquad |X|\,, \qquad X^{\pm}\,.$$

2.5.1 Simple Functions

When $X : \Omega \to \mathbb{R}$ takes only a finite number of values, we say that X is a *simple function.*

The prime example of a simple function is the *indicator function* $X = \mathbf{1}_A$ for some set $A \in \mathcal{F}$ (to wit, $\mathbf{1}_A(\omega) := 1$ if $\omega \in A$, and $\mathbf{1}_A(\omega) := 0$ if $\omega \in \Omega \setminus A$).

More generally, a simple function $X : \Omega \to \mathbb{R}$ has range $X(\Omega) = \{f_1, \cdots, f_K\}$, where $f_1, \cdots, f_K$ are distinct real numbers and K an integer. Then the sets

$$A_k := \{\omega \in \Omega : X(\omega) = f_k\} = X^{-1}(\{f_k\})\,, \qquad k = 1, \cdots, K$$

[11] Such constructions are highly non-trivial, and require great ingenuity and care. In a sense, these difficulties reflect the price one has to pay, in order to obtain measures that are *countably* – as opposed to merely finitely – *additive*.

are disjoint, their union is Ω (i.e., they form a "partition" of the sample space), and the simple function can be expressed as a superposition of indicators: $X = \sum_{k=1}^{K} \xi_k \, \mathbf{1}_{A_k}$.

We shall denote by $\mathcal{S}$ the class of all simple functions, and by $\mathcal{S}_+$ its subclass of nonnegative simple functions. We shall also consider $\mathcal{S}_+^*$, the class of functions $X : \Omega \to \Xi$ with $\Xi := \{\xi_1, \cdots, \xi_K\} \subset [0, \infty]$ a finite set.

As we shall see in Proposition 4.1 below, *every nonnegative measurable function is the pointwise, increasing limit of a sequence of nonnegative, simple functions.* This property will be of the utmost importance in the study of the LEBESGUE integral.

2.6 Random Variables

Let us fix a **probability space** $(\Omega, \mathcal{F}, \mathbb{P})$. On such a space, a measurable function $X : \Omega \to \mathbb{R}$ will be called *Random Variable*. [12] We think then of $X(\omega)$ as a numerical characteristic associated with the random experiment, whose realization $\omega \in \Omega$ has been observed. [13]

For instance, in Example **(a)** above, for any given $\omega = (\omega_1, \cdots, \omega_n) \in \{0, 1\}^n$ the quantity $X(\omega) = \sum_{i=1}^{n} \omega_i$ is the total number of heads obtained in the n coin tosses. Think also of the mass of a widget coming off an assembly line; or of the lifetime of an infant at birth.

The *distribution* of the random variable X is the probability measure induced by X on the BOREL subsets of $\mathbb{R}$, namely

$$\mu_X(B) := \mathbb{P}\big(\{\omega \in \Omega : X(\omega) \in B\}\big) \equiv \mathbb{P}(X \in B) = \big(\mathbb{P} \circ X^{-1}\big)(B) \tag{2.3}$$

for $B \in \mathcal{B}(\mathbb{R})$. This probability measure generates the *probability distribution function*

$$F_X(x) := \mu_X\big((-\infty, x]\big) = \mathbb{P}(X \le x)\,, \qquad x \in \mathbb{R} \tag{2.4}$$

of the random variable X.

It is clear that this function $F_X : \mathbb{R} \to [0, 1]$ is nondecreasing; it is also not hard to see that it is right-continuous and $F(-\infty) = 0$, $F(\infty) = 1$ (exercise).

We say that two random variables X and Y are *identically distributed,* if $\mu_X \equiv \mu_Y$, or equivalently if $\mathbb{P}(X \le x) = \mathbb{P}(Y \le x)$ holds for every $x \in \mathbb{R}$.

[12] The terminology is certainly very strange, as in: "What do you mean, is this a function or is it a variable?". We must apologize for it in advance. The reason for this weird terminology is to be found in the long time, the centuries, during which the theory of Probability developed its own methods and terminology, distinct from Mathematical Analysis; as well as in the fact that rather late in the game, in the first part of the 20^{th} century, the realization dawned that the two were just different facets of the same thing.

[13] In this case, the measurability requirement simply means that we can assign a probability to the event $X^{-1}(I) = \{\omega \in \Omega \,|\, X(\omega) \in I\}$ that the numerical characteristic associated with our experiment take values in any given interval $I \subset \mathbb{R}$; in other words, $X^{-1}(I) \in \mathcal{F}$. This is a minimal requirement on the function $X : \Omega \to \mathbb{R}$; without it, we cannot even have a model, let alone cannot start to develop a theory of Probability that makes allowances for observable, numerical characteristics of random experiments.

Caution: *Two very different random variables can be identically distributed.*

Indeed, take $(\Omega, \mathcal{F}) \equiv ([0,1], \mathcal{B}([0,1]))$ with Lebesgue measure $\mathbb{P} = \lambda|_{[0,1]}$ and consider the mappings $X(\omega) = \omega$, $Y(\omega) = 1 - \omega$, $\omega \in \Omega$. It is clear that these two mappings have the same distribution

$$F_X(x) \,=\, F_Y(x) \,=\, (x \wedge 1) \vee 0\,, \qquad x \in \mathbb{R}\,,$$

the so-called "uniform probability distribution" function on $[0,1]$.

• In a completely analogous manner, we call an $\mathcal{F}-$measurable $\mathfrak{X} : \Omega \to \mathbb{R}^n$ a *Random Vector,* and the measure $\mu_{\mathfrak{X}}$ induced on $\mathcal{B}(\mathbb{R}^n)$ as in (2.3) the *distribution* of this random vector $\mathfrak{X} = (X_1, \cdots, X_n)$.

With $(-\infty, \mathbf{x}] = \{y \in \mathbb{R}^n \,|\, y_j \leq x_j\,,\ \forall\ j = 1, \cdots, n\}$ and $\{\mathfrak{X} \leq \mathbf{x}\} = \bigcap_{j=1}^n \{X_j \leq x_j\}$ for $\mathbf{x} = (x_1, \cdots, x_n) \in \mathbb{R}^n$, we define by analogy with (2.4) the *probability distribution function* of the random vector $\mathfrak{X}$ as

$$F_{\mathfrak{X}}(\mathbf{x}) \,=\, \mathbb{P}(\mathfrak{X} \leq \mathbf{x})\,, \qquad \mathbf{x} \in \mathbb{R}^d\,.$$

Definition 2.4. *Consider a random variable $X : \Omega \to \mathbb{R}$ on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and recall the probability measure μ_X induced on the* BOREL *sets of the real line. We say that X has*

(i) Discrete Distribution, *if*

$$\mu_X \,=\, \sum_{j \in \mathcal{J}} \mathfrak{p}_j\, \delta_{\xi_j}$$

for some (at most) countable set $\{\xi_j\}_{j \in \mathcal{J}}$ and numbers $\mathfrak{p}_j > 0$, $j \in \mathcal{J}$ with $\sum_{j \in \mathcal{J}} \mathfrak{p}_j = 1$;

(ii) Diffuse Distribution, *if there exists a* BOREL *measurable function $f_X : \mathbb{R} \to [0, \infty)$ such that $\mu_X(B) = \int_B f_X(\xi)\, d\xi$ holds for every $B \in \mathcal{B}(\mathbb{R})$. We call this $f_X(\cdot)$ the* probability density function *of the random variable X.*

In particular, the probability distribution function

$$F_X(x) := \mu_X\big((-\infty, x]\big) = \int_{-\infty}^{x} f_X(\xi)\, d\xi \,= \mathbb{P}(X \leq x)\,, \qquad x \in \mathbb{R}$$

of a random variable with a diffuse distribution is continuous.

Caution: There exist random variables X whose distribution μ_X is not diffuse, yet $F_X(\cdot) = \mu_X\big((-\infty, \cdot\,]\big)$ is continuous. For an example, see Remark 3.2 below, or chapter 15. The term "absolutely continuous" is also used in place of "diffuse".

We shall study several probability distribution functions in the next chapters.

Remark 2.5. *Nondecreasing Functions of a Real Variable:* Let us recall (for instance, from NATANSON (1955); CHUNG (1974), chapter 1; or BILLINGSLEY (1986), section 31) some basic facts concerning a nonconstant, nondecreasing function $F : \mathbb{R} \to \mathbb{R}$.

For such a function, the limits

$$F(\infty) := \lim_{x \to \infty} F(x) \qquad \text{and} \qquad F(-\infty) := \lim_{x \to \infty} F(-x)$$

exist; so do, at each $x \in \mathbb{R}$, the limits

$$F(x+) := \lim_{h \downarrow 0} F(x+h) \;\geq\; F(x) \;\geq\; F(x-) := \lim_{h \downarrow 0} F(x-h)\,.$$

We say that $F(\cdot)$ has a *jump* at $x \in \mathbb{R}$ if $\Delta F(x) := F(x+) - F(x-) > 0$; and in this case $\Delta F(x)$ is called the *size of the jump*. Jumps are the only possible discontinuities for $F(\cdot)$, and there can be (at most) countably many of them.

Such a function is differentiable except on a set of zero LEBESGUE measure; the derivative $F'(\cdot) \geq 0$ is measurable, and satisfies

$$\int_a^b F'(x)\,dx \;\leq\; F(b) - F(a)\,, \qquad \forall\; a < b\,. \tag{2.5}$$

This inequality can be strict; again, see Remark 3.2 or chapter 15 for examples. On the other hand, if $F(-\infty) = 0$, $F(\infty) < \infty$ and there exists an integrable function $f : \mathbb{R} \to [0,\infty)$ with $F(x) = \int_{-\infty}^{x} f(\xi)\,d\xi$, $\forall\, x \in \mathbb{R}$, then $F'(\cdot) = f(\cdot)$ holds outside a set of zero LEBESGUE measure, and (2.5) holds with equality.

Exercise 2.5. Construct a distribution function which is discontinuous at every rational point, and continuous at all irrational points on the real line.

Conversely: is there a distribution function which is discontinuous at every irrational point, and continuous at all rational points on the real line?

2.7 Integral

For a nonnegative, simple function $X \in \mathcal{S}_+^*$ on an **arbitrary measure space** $(\Omega, \mathcal{F}, \mathbb{P})$, its LEBESGUE *Integral* is defined as

$$\mathbb{E}(X) \equiv \int_\Omega X\,\mathrm{d}\mathbb{P} := \sum_{k=1}^{K} \xi_k \cdot \mathbb{P}\big(X^{-1}(\{\xi_k\})\big) = \sum_{k=1}^{K} \xi_k \cdot \mathbb{P}\big(X = \xi_k\big). \tag{2.6}$$

Here $X(\Omega) := \{\xi_1, \cdots, \xi_K\} \subset [0,\infty]$ is the range of X, and we adopt the convention $\infty \cdot 0 = 0$ and the shorthand notation $\{X = \xi\} = \{\omega \in \Omega : X(\omega) = \xi\}$. We have clearly

$$\mathbb{E}(cX) = c\,\mathbb{E}(X) \qquad \text{for any real constant } c \geq 0 \text{ and any } X \in \mathcal{S}_+\,, \tag{2.7}$$

as well as $\mathbb{E}(\alpha X + \beta Y) = \alpha \mathbb{E}(X) + \beta\,\mathbb{E}(Y)$, or equivalently

$$\int_\Omega \big(\alpha X + \beta Y\big)\,\mathrm{d}\mathbb{P} \;=\; \alpha \int_\Omega X\,\mathrm{d}\mathbb{P} + \beta \int_\Omega Y\,\mathrm{d}\mathbb{P}$$

for any two simple functions $X \in \mathcal{S}_+^*$, $Y \in \mathcal{S}_+^*$ and real constants $\alpha \geq 0$, $\beta \geq 0$ (argue this out carefully).

Example 2.1. On the unit interval $\Omega = [0,1]$ with the σ-algebra $\mathcal{F}$ of its BOREL sets and $\mathbb{P} = \lambda =$ LEBESGUE measure, consider the simple function $X = \mathbf{1}_{\mathbb{Q}\cap[0,1]}$ which takes the value 1 on the rationals of the unit interval, and the value 0 on its irrationals. This function is *not* RIEMANN integrable.

But its LEBESGUE integral not only exists, it is the easiest thing in the world to find: $\mathbb{E}(X) = 1 \cdot \lambda(\mathbb{Q} \cap [0,1]) = 0$, as the set of rationals has zero LEBESGUE measure.

• We define the LEBESGUE Integral of an arbitrary measurable function $X : \Omega \to [0, \infty)$ as

$$\mathbb{E}(X) \equiv \int_\Omega X \, \mathrm{d}\mathbb{P} := \sup_{Y \in \mathcal{S},\ 0 \le Y \le X} \int_\Omega Y \, \mathrm{d}\mathbb{P} \,. \tag{2.8}$$

Here the supremum [14] on the right-hand side is taken over all non-negative simple functions $Y \in \mathcal{S}_+$ satisfying $Y \le X$ *pointwise*, that is, $Y(\omega) \le X(\omega)$ for every $\omega \in \Omega$.

• A measurable function $X : \Omega \to [0, \infty)$ is called *integrable*, if the supremum of (2.8) is finite; to wit, if $\mathbb{E}(X) < \infty$.

For any given set $B \in \mathcal{F}$, we denote by $\int_B X \, \mathrm{d}\mathbb{P}$ the integral

$$\int_\Omega X \, \mathbf{1}_B \, \mathrm{d}\mathbb{P} \equiv \mathbb{E}(X \mathbf{1}_B).$$

It can be checked easily that the set function $\mathcal{F} \ni B \longmapsto \nu(B) := \int_B X \, \mathrm{d}\mathbb{P} \in [0, \infty]$ is a measure, and has the property

$$\mathbb{P}(B) = 0 \implies \nu(B) = 0.$$

Nota Bene: It is very important to note that *this definition of the integral* – unlike that of the more familiar RIEMANN integral – *assumes no topological structure on the part of the space* Ω *whatsoever.*

Remark 2.6. The definition in (2.8) is an instance of the "method of exhaustion", invented in antiquity by EUDOXOS (4^{th} century BC) and greatly developed by ARCHIMEDES (3^{rd} century BC) for the purpose of calculating areas and volumes of geometric figures. [15] This method is the precursor of the modern concept of *limit:* according to it, a convex region is approximated by inscribed (or circumscribed) polygons, whose areas are relatively easy to calculate, and then the number of vertices of the polygons is increased until the area of the region has been "exhausted".

The much later work of NEWTON and LEIBNIZ in the late 17^{th} century made this method a systematic and powerful tool for such calculations. By the twentieth century the main applications of this theory had shifted from geometry and elementary mechanics to differential equations, convergence of orthogonal expansions, and Probability Theory. □

We have clearly for this integral the *monotonicity property*

$$\mathbb{E}(X_1) \le \mathbb{E}(X_2)\,, \quad \text{if} \quad 0 \le X_1 \le X_2\,, \tag{2.9}$$

as well as (argue this one out in detail!) the homogeneity property

$$\mathbb{E}(cX) = c\,\mathbb{E}(X) \quad \text{for any } c \in [0, \infty) \text{ and any measurable } X : \Omega \to [0, \infty) \tag{2.10}$$

[14] The *supremum* (or "least upper bound") $\sup(A)$ of a set A of real numbers, is the smallest element of $\mathbb{R} \cup \{+\infty\}$ that is greater than or equal to every number in A. The *infimum* (or "greatest lower bound") $\inf(A)$ of A is the greatest element of $\mathbb{R} \cup \{-\infty\}$ that is less than or equal to every number in A.

[15] As L.C. YOUNG (1928) puts it: "The Theory of Integration ... has long been one of the most useful tools in Mathematics. Its methods were already employed with success by the ancient Greeks, in their investigations about Areas and Volumes. They possessed the method of exhaustion, the method of series. They were very clear about the idea of limit and this perhaps made them suspicious of the unsound method of infinitesimals, as results thus obtained were always established independently. After the Dark Ages the rediscovery of this last method and the use of the symbolism of Algebra rendered possible the creation of the Calculus by NEWTON and LEIBNITZ."

• Now for an arbitrary measurable function $X : \Omega \to \mathbb{R}$, its positive and negative parts $X^{\pm} = \max(\pm X, 0)$, as well as its absolute value $|X| = X^+ + X^-$, are all measurable; recall Exercise 2.4.

If at least one of the quantities $\mathbb{E}(X^+)$, $\mathbb{E}(X^-)$ is finite, we define the LEBESGUE integral of the real-valued function X as

$$\mathbb{E}(X) \equiv \int_\Omega X \, d\mathbb{P} := \int_\Omega X^+ \, d\mathbb{P} - \int_\Omega X^- \, d\mathbb{P} = \mathbb{E}(X^+) - \mathbb{E}(X^-) . \tag{2.11}$$

We say that X is *integrable*, if $|X|$ is integrable; in this case $X^{\pm}$ are also clearly integrable (and also the other way around), on account of (2.15). It is seen from this definition that

$$|\,\mathbb{E}(X)\,| \le \mathbb{E}\big(|X|\big) , \tag{2.12}$$

then from (2.10) that we have the homogeneity property

$$\mathbb{E}(cX) = c\,\mathbb{E}(X) \qquad \text{for any real constant } c, \text{ and any integrable } X : \Omega \to \mathbb{R} . \tag{2.13}$$

We also have the following linearity property, which will be proved in section 4.3.

Proposition 2.1. Linearity of the Integral: *For any real constants α, β and any two integrable functions X and W, the function $\alpha X + \beta W$ is also integrable and we have the linearity property $\mathbb{E}(\alpha X + \beta W) = \alpha \mathbb{E}(X) + \beta\,\mathbb{E}(W)$, or equivalently*

$$\int_\Omega \big(\alpha X + \beta W\big)\, d\mathbb{P} = \alpha \int_\Omega X \, d\mathbb{P} + \beta \int_\Omega W \, d\mathbb{P} . \tag{2.14}$$

Exercise 2.6. Establish carefully the homogeneity properties (2.10) and (2.13), as well as (2.12).

In particular, it follows from Proposition 2.1 that

$$\mathbb{E}\big(|X|\big) = \mathbb{E}(X^+) + \mathbb{E}(X^-) . \tag{2.15}$$

Thus, $|X|$ is integrable if, and only if, both $X^{\pm}$ are integrable.

2.8 Expectation

When $\mathbb{P}(\Omega) = 1$, the quantity of (2.6) is simply the center of gravity

$$\frac{\sum_{k=1}^K \xi_k \, p_k}{\sum_{k=1}^K p_k}$$

for a distribution of mass that assigns weight $p_k := \mathbb{P}(X^{-1}(\{\xi_k\}))$ to the site ξ_k, for each $k = 1, \cdots, K$; it is called the *Expectation* of the simple random variable X and we denote it

$$\boxed{\mathbb{E}(X) \equiv \sum_{k=1}^K \xi_k \cdot \mathbb{P}\big(X = \xi_k\big) .}$$

Then we proceed with the definition of the expectation for more general random variables as in (2.8), (2.11). In particular, we see that two random variables X and Y are identically distributed, if and only if

$$\mathbb{E}\big[\Psi(X)\big] \,=\, \mathbb{E}\big[\Psi(Y)\big] \tag{2.16}$$

holds for every indicator function $\Psi = \mathbf{1}_{(-\infty,x]}$ and $x \in \mathbb{R}$.

Exercise 2.7. For a random variable $X : \Omega \to \mathbb{N}_0$ which takes nonnegative integer values, show:

$$\mathbb{E}(X) \,=\, \sum_{k\in\mathbb{N}} \mathbb{P}\big(X \geq k\big)\,.$$

2.9 Definition of Independence

For a fixed event F in a **probability space** $(\Omega, \mathcal{F}, \mathbb{P})$ with $\mathbb{P}(F) > 0$, we define the *conditional probability* measure *given* F by

$$\boxed{\mathbb{P}_F(E) \,\equiv\, \mathbb{P}(E\,|\,F) \,:=\, \frac{\mathbb{P}(E\cap F)}{\mathbb{P}(F)}\,,\quad E\in\mathcal{F}\,.} \tag{2.17}$$

Now suppose that for some event $E \in \mathcal{F}$ we have $\mathbb{P}_F(E) = \mathbb{P}(E)$, i.e., that knowledge about the occurrence (or not) of the event F does not change the probability assigned by the measure $\mathbb{P}$ to the event E. Or equivalently, that

$$\boxed{\mathbb{P}(E\cap F) \,=\, \mathbb{P}(E)\,\mathbb{P}(F)\,,} \tag{2.18}$$

a relation which is symmetric in E and F and unambiguous, even when the probabilities $\mathbb{P}(E)$ or $\mathbb{P}(F)$ vanish.

We say that the two events $E,\ F$ are **independent,** if (2.18) holds. It is interesting to check that, if this is the case, then $E,\ F^c$ (and $E^c,\ F$ as well as $E^c,\ F^c$), are also independent. [16]

Similarly, we say that two random variables $X : \Omega \to \mathbb{R}$ and $Y : \Omega \to \mathbb{R}$ are **independent,** if $X^{-1}(A)$ and $Y^{-1}(B)$ are independent events for any BOREL subsets A and B of the real line (equivalently, if $X^{-1}\big((-\infty, a]\big)$ and $Y^{-1}\big((-\infty, a]\big)$ are independent events for any reals $a,\ b$.)

Suppose, for instance, that we throw a die twice, so $\Omega = \{\,(\omega_1,\omega_2) \,:\, 1 \leq \omega_1,\omega_2 \leq 6\,\}$ and consider the events $A = \{\,(\omega_1,\omega_2) \,:\, \omega_1+\omega_2 = 7\}$, $B = \{\,(\omega_1,\omega_2) \,:\, \omega_1+\omega_2 = 6\}$ and $C = \{\,(\omega_1,\omega_2) \,:\, \omega_1 = 3\}$. Then we have $\mathbb{P}(A) = \mathbb{P}(C) = 6/36 = 1/6$ and $\mathbb{P}(A\cap C) = \mathbb{P}(B\cap C) = 1/36$, whereas $\mathbb{P}(B) = 5/36$. Thus, the events A and C are independent, but the events B and C are not.

[16] It seems that this notion was introduced, in an intuitive manner and as a computational tool, by CARDANO.

Definition 2.5. Independent Events, Random Variables. *Let us place ourselves on a* **probability space** $(\Omega, \mathcal{F}, \mathbb{P})$.

(i) The events in an arbitrary family $\mathcal{E} = \{E_\alpha\}_{\alpha \in A}$ *are said to be* independent, *if for any* $n \in \mathbb{N}$ *and* $\{\alpha_1, \cdots, \alpha_n\} \subseteq A$ *we have*

$$\mathbb{P}\left(\bigcap_{j=1}^{n} E_{\alpha_j}\right) = \prod_{j=1}^{n} \mathbb{P}(E_{\alpha_j}).$$

(ii) The random variables in a family $\{X_\alpha\}_{\alpha \in A}$ *are said to be* independent, *if the events* $\{X_\alpha^{-1}(B_\alpha)\}_{\alpha \in A}$ *are independent for given any family of Borel subsets* $\{B_\alpha\}_{\alpha \in A} \subseteq \mathcal{B}(\mathbb{R})$. *In other words, if*

$$\mathbb{P}\big(\{X_{\alpha_1} \in B_{\alpha_1}\} \cap \cdots \cap \{X_{\alpha_n} \in B_{\alpha_n}\}\big) = \prod_{j=1}^{n} \mathbb{P}\big(\{X_{\alpha_j} \in B_{\alpha_j}\}\big)$$

holds for any $n \in \mathbb{N}$, $\{\alpha_1, \cdots, \alpha_n\} \subseteq A$, *and* BOREL *subsets* $B_{\alpha_1}, B_{\alpha_2}, \cdots, B_{\alpha_n}$ *of the real line.*

(iii) Two collections $\mathcal{G}$ *and* $\mathcal{H}$ *of events are called independent, if for every* $G \in \mathcal{G}$ *and* $H \in \mathcal{H}$ *we have*

$$\mathbb{P}(G \cap H) = \mathbb{P}(G) \cdot \mathbb{P}(H).$$

(iv) More generally, suppose $\mathcal{E}^{(1)}, \mathcal{E}^{(2)}, \cdots$ *are collections of events; we say that these collections are independent, if for any* $n \in \mathbb{N}$, *any distinct integers* $i_1, \cdots, i_n$, *and any events* $E_j \in \mathcal{E}^{(j)}$, $j \in \{i_1, \cdots, i_n\}$ *we have*

$$\mathbb{P}\Big(E_{i_1} \bigcap \cdots \bigcap E_{i_n}\Big) = \prod_{k=1}^{n} \mathbb{P}(E_{i_k}).$$

It is easy to verify that, if the events of an arbitrary family $\{E_\alpha\}_{\alpha \in \mathcal{I}}$ are independent, then $\{F_\alpha\}_{\alpha \in \mathcal{I}}$ are also independent, where F_α can be either E_α or E_α^c.

Similarly: *if* $\{X_\alpha\}_{\alpha \in \mathcal{I}}$ *are independent random variables and* $\{f_\alpha\}_{\alpha \in \mathcal{I}}$ *are* BOREL *measurable functions, then* $\{f_\alpha(X_\alpha)\}_{\alpha \in \mathcal{I}}$ *are also independent.* Indeed, if $\{\alpha_j\}_{j=1}^n \subseteq \mathcal{I}$ is any set of n indices, and $\{B_j\}_{j=1}^n$ are BOREL subsets of $\mathbb{R}$, then

$$\mathbb{P}\left[\bigcap_{j=1}^{n} \left\{f_{\alpha_j}(X_{\alpha_j}) \in B_j\right\}\right] = \mathbb{P}\left[\bigcap_{j=1}^{n} \left\{X_{\alpha_j} \in f_{\alpha_j}^{-1}(B_j)\right\}\right]$$

$$= \prod_{j=1}^{n} \mathbb{P}\left(X_{\alpha_j} \in f_{\alpha_j}^{-1}(B_j)\right) = \prod_{j=1}^{n} \mathbb{P}\left[\, f_{\alpha_j}(X_{\alpha_j}) \in B_j \,\right].$$

Exercise 2.8. If $\{X_\alpha\}_{\alpha \in A}$ is a family of independent random variables, then the $\sigma-$algebras generated by disjoint subfamilies are independent.

Caution: *Three events can be pairwise independent, without being independent.*

To see this, consider a tetrahedron with faces $\omega_1, \cdots, \omega_4$: the first is painted red, the second blue, the third green, while the forth contains all three of these colors. We toss the tetrahedron, and it lands on one of its faces (all possibilities equally likely). Consider the events $R = \{\omega_1, \omega_4\}$, $B = \{\omega_2, \omega_4\}$, $G = \{\omega_3, \omega_4\}$ that the face on which the tetrahedron lands contains the color red, the color blue, and the color green, respectively.

Assigning normalized counting measure to all the subsets of $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4\}$, we have $\mathbb{P}(R) = \mathbb{P}(B) = \mathbb{P}(G) = 2/4$ and $\mathbb{P}(R \cap B) = \mathbb{P}(B \cap G) = \mathbb{P}(G \cap R) = 1/4$, so the events are independent *pairwise* (that is, when considered two by two); however, they are *not* independent, because $\mathbb{P}(R \cap B \cap G) = 1/4 \neq 1/8 = \mathbb{P}(R)\,\mathbb{P}(B)\,\mathbb{P}(G)$. □

The preceding is just a listing of basic definitions. We shall study the notion and the properties of independence in some detail throughout this course, so we shall keep returning to these notions again and again in subsequent chapters and sections.

3 Some Examples

We provide in this chapter some examples of random variables and their distributions, as they arise naturally in specific contexts.

3.1 The Magic of Coincidences, and a Foretaste of POISSON

We have invited n gentlemen to a major black-tie, top-hat gala event (cf. Figure 1). Each of them throws his hat at random in the center of the room. The hats are then mixed up (shuffled) thoroughly, then at the end of the event each gentleman is assigned, at random, a hat. What is the probability that at least one gentleman is assigned his own hat?

Or, to put this question slightly differently, in an open-ended fashion: *How likely are coincidences?*

The sample space here is the symmetric group $\Omega = \Sigma_n$, the set consisting of all $|\Omega| = n!$ permutations $\omega = \big(\omega(1), \cdots, \omega(n)\big)$ of the integers $\{1, \cdots, n\}$. We consider normalized counting measure $\mathbb{P}(A) = |A|/n!$ on the $\sigma-$algebra $\mathcal{F}$ of all subsets $A \subseteq \Omega$ of this space; in particular, "all permutations are considered equally likely". We are interested in the event

$$A^\star = \bigcup_{j=1}^{n} A_j\,, \qquad A_j := \{\omega \in \Omega : \omega(j) = j\}\,, \tag{3.1}$$

for which the inclusion-exclusion formula gives

$$\mathbb{P}(A^\star) = \sum_{\mathcal{J} \subseteq \{1,\cdots,n\},\ \mathcal{J} \neq \emptyset} (-1)^{|\mathcal{J}|+1}\ \mathbb{P}\Big(\bigcap_{j \in \mathcal{J}} A_j\Big)\,.$$

Since in our present context

$$\mathbb{P}\Big(\bigcap_{j \in \mathcal{J}} A_j\Big) = \frac{\big|\{\omega : \omega(j) = j,\ \forall\, j \in \mathcal{J}\}\big|}{n!} = \frac{(n - |\mathcal{J}|)!}{n!}\,,$$

we deduce

$$\mathbb{P}(A^\star) = \sum_{k=1}^{n} (-1)^{k+1}\, \frac{(n-k)!}{n!} \cdot \Big|\{\mathcal{J} \subseteq \{1, \cdots, n\} : |\mathcal{J}| = k\}\Big|$$

$$= \sum_{k=1}^{n} (-1)^{k+1}\, \frac{(n-k)!}{n!} \cdot \binom{n}{k} = \sum_{k=1}^{n} \frac{(-1)^{k+1}}{k!} = 1 - \sum_{k=0}^{n} \frac{(-1)^k}{k!}\,.$$

In particular, the probability that *"no gentleman receives his own hat"* is

$$\sum_{k=0}^{n} \frac{(-1)^k}{k!}\,.$$

Figure 1: Inauguration Day, January 1961 (the last one featuring top hats).

We have used some very basic facts from combinatorial analysis; cf. Appendix, Chapter 13. In particular, that there are

$$\binom{n}{k} = \frac{n!}{k!\,(n-k)!}$$

ways to select k objects out of n, without replacement and without regard to order.

As $n \to \infty$, this expression tends to $(1/e)$, the asymptotic probability of "no coincidence"; whereas $\mathbb{P}(A^\star)$ tends to $1 - (1/e) \simeq .63$; the so-called "asymptotic probability of (at least one) coincidence".

Exercise 3.1. The ST. BASIL of Caesarea [17] problem; and a foretaste of POISSON: In the same setting as above, what is the probability that *exactly one* gentleman is assigned his own hat?

More generally, show that the probability that *exactly* m gentlemen are assigned their own hats, is given by

$$\frac{1}{m\,!} \sum_{k=0}^{n-m} \frac{(-1)^k}{k\,!}\,, \qquad m = 0, 1, \cdots, n\,.$$

Observe that these probabilities converge, as $n \to \infty$, to the weights

$$e^{-1}\,\frac{1}{m!}\,, \qquad m = 0, 1, \cdots$$

of the so-called "POISSON distribution with parameter $\lambda = 1$" (see subsection 3.3 below).

Exercise 3.2. You have invited ten married couples to your house for dinner. Exasperated with the subtleties of seating arrangements, you decide to rely on a random draw. What is the probability that at least one wife will be seated next to her husband?

3.2 BERNOULLI and Binomial Distributions

Let us recall the space $\Omega = \{0,1\}^n$ for n *coin tosses;* it has $N = 2^n$ elements of the form $\omega = (\omega_1, \cdots, \omega_n)$ with $\omega_i \in \{0,1\}$. Each ω_i is interpreted as the outcome of the i-th coin toss, with $\omega_i = 1$ for "heads" and $\omega_i = 0$ for "tails". We endow this space with the collection $\mathcal{F}$ of all its subsets.

[17] A Father and pillar of the early Christian Church, ST. BASIL was Bishop of Caesarea in Asia Minor in the early 4^{th} century. He was a scion of a very wealthy family, but gave all his possessions to the poor and devoted himself to rigorous monastic life, to learning (he had an unparalleled knowledge and understanding of the Classics) and to writing (his epistles and his theological works are also literary masterpieces). Tradition has it that a governor of the region, bent on extracting as much tithe from the population as he possibly could, ordered all citizens, rich and poor alike but mostly poor, to turn in all their valuables. The good Bishop intervened on behalf of his flock, and somehow convinced the governor to return the valuables; but to whom should each item be returned? The Bishop asked the local bakers to prepare many loafs of bread; hid the valuables in them; then asked the governor's minions to distribute the loafs to the various households. Miracle of miracles: each household received exactly the valuables it had surrendered.

Suppose we attach to this $(\Omega, \mathcal{F})$ the normalized counting measure, or "uniform distribution", $\mathbb{P}$ with $\mathfrak{p}_\omega = 1/|\Omega| = 2^{-n}$ for each $\omega \in \Omega$. Then every subset A of Ω is assigned probability $\mathbb{P}(A) = 2^{-n}\,|A|$. We consider the coördinate mappings ("random variables") $X_i(\omega) = \omega_i$, $i = 1, \cdots, n$, as well as the total number of "heads"

$$S_n(\omega) \,:=\, \sum_{i=1}^{n} X_i(\omega) \,=\, \sum_{i=1}^{n} \omega_i\,, \qquad \omega \in \Omega\,.$$

What is the distribution of this random variable? Let us fix $k \in \{0, 1, \cdots, n\}$ and observe that

$$2^n \cdot \mathbb{P}(S_n = k) \,=\, \Big|\, \Big\{\, \omega \in \Omega \;\Big|\; \sum_{j=1}^{n} \omega_j = k \,\Big\} \Big| \,=\, \binom{n}{k} \,=\, \frac{n!}{k!\,(n-k)!}\,,$$

the binomial coëfficient that measures in how many ways we can choose k heads (thus $n-k$ tails) out of a total of n throws, without regard to order. Then every subset A of Ω is assigned probability $\mathbb{P}(A) = 2^{-n}|A|$.

We say that S_n has the "Binomial distribution" with parameters $(n; 1/2)$ and write $S_n \sim \mathrm{Bin}\,(n; 1/2)$. The parameter $p = 1/2$ refers to the "success probability" that one expects of a *fair coin*.

Now suppose we are dealing with an "unfair coin", so for some given $p \in (0,1)$ we set

$$\mathfrak{p}_\omega \,:=\, p^{\,\#\,\text{of}\;1\text{'s in}\;\omega}\,(1-p)^{\,\#\,\text{of}\;0\text{'s in}\;\omega} \,=\, p^{\,S_n(\omega)}\,(1-p)^{\,n-S_n(\omega)}\,, \quad \omega \in \Omega \tag{3.2}$$

and

$$\mathbb{P}(A) \,:=\, \sum_{\omega \in A} \mathfrak{p}_\omega \qquad \text{for any} \quad A \subset \Omega\,. \tag{3.3}$$

This is a *bona fide* attribution, since the binomial theorem guarantees that we have

$$\sum_{\omega \in \Omega} \mathfrak{p}_\omega = \sum_{k=0}^{n} \Big|\, \{\, \omega \in \Omega \;:\; \sum_{j=1}^{n} \omega_j = k \,\} \Big|\, p^k\,(1-p)^{n-k} = \sum_{k=0}^{n} \binom{n}{k}\, p^k\,(1-p)^{n-k} \,=\, 1\,;$$

and it reduces to the previous attribution of the normalized counting measure, when $p = 1/2$. We stress that the attribution is otherwise arbitrary at this point; we shall discuss the reasoning behind this particular attribution below. In "defense" of (3.2), let us just say at this point that it agrees with the attribution $\mathfrak{p}_\omega = 2^{-n}$ made earlier, when $p = 1/2$.

As a small sanity check, let us fix $i \in \{1, \cdots, n\}$, and observe from (3.3) and (3.2) that, with $\Omega_k^* := \{\omega \in \Omega \,:\, \omega_i = 1,\, \sum_{j \neq i} \omega_j = k\}$, we have

$$\begin{aligned} \mathbb{P}(X_i = 1) \,&=\, \sum_{\omega \in \Omega\,:\,\omega_i = 1} \mathfrak{p}_\omega \,=\, \sum_{k=0}^{n-1} \left(\sum_{\omega \in \Omega_k^*} p^{k+1}(1-p)^{n-1-k} \right) \\ &=\, p \sum_{k=0}^{n-1} p^k (1-p)^{n-1-k} \cdot \Big|\, \Omega_k^* \Big| \,=\, p \sum_{k=0}^{n-1} p^k (1-p)^{n-1-k} \cdot \binom{n-1}{k}\,. \end{aligned}$$

This is because, in order to compute the cardinality of Ω_k^*, we need to find the number of possibilities in which k objects can be chosen out of $n-1$, without replacement and without regard to order. This reduces to

$$\mathbb{P}(X_i = 1) \,=\, p \sum_{k=0}^{n-1} p^k (1-p)^{(n-1)-k} \cdot \binom{n-1}{k} \,=\, p \cdot \big(p + (1-p)\big)^{n-1} \,=\, p$$

by the binomial theorem. Similarly, $\mathbb{P}(X_i = 0) = 1 - p$, so each random variable X_i, $i = 1, \cdots, n$ has the so-called **BERNOULLI (1713) distribution**:

$$\mathbb{P}(X_i = 1) \,=\, p\,, \qquad \mathbb{P}(X_i = 0) = 1 - p\,. \tag{3.4}$$

This, of course, was to be expected! In a similar fashion, just with slightly more elaborate computations, for any two given integers $1 \le i \ne j \le n$ one can obtain

$$\mathbb{P}\big(X_i = \vartheta,\, X_j = \varrho\big) \,=\, p(1-p) \,=\, \mathbb{P}\big(X_i = \vartheta\big) \cdot \mathbb{P}\big(X_j = \varrho\big)$$

for $(\vartheta,\, \varrho) = (0,1)$ and $(\vartheta,\, \varrho) = (1,0)$; as well as

$$\mathbb{P}\big(X_i = 1,\, X_j = 1\big) \,=\, p^2 \,=\, \mathbb{P}\big(X_i = 1\big) \cdot \mathbb{P}\big(X_j = 1\big)\,,$$

$$\mathbb{P}\big(X_i = 0,\, X_j = 0\big) = (1-p)^2 = \mathbb{P}\big(X_i = 0\big) \cdot \mathbb{P}\big(X_j = 0\big)\,.$$

In other words, with this specification of probability measure, the random variables X_i and X_j are independent.

Even more to the point, for any given $x = (x_1, \cdots, x_n) \in \Omega = \{0,1\}^n$, we can compute this way

$$\mathbb{P}\big(X_1 = x_1, \cdots, X_n = x_n\big) \,=\, p^{\,\sum_{i=1}^n x_i}\, (1-p)^{\,n - \sum_{i=1}^n x_i}\,. \tag{3.5}$$

Extending this argument one step further, we deduce that *the random variables* $X_1, X_2, \cdots, X_n$ *are independent:* for any $\mathcal{J} = \{j_1, \cdots, j_k\} \subseteq \{1, \cdots, n\}$ and any $(x_1, \cdots, x_k) \in \{0,1\}^k$, we have

$$\mathbb{P}\big(X_{j_1} = x_1, \cdots, X_{j_k} = x_n\big) \,=\, p^{\,\sum_{i=1}^k x_i}\, (1-p)^{\,k - \sum_{i=1}^k x_i}\,.$$

It is now clear, perhaps, why we made the particular attribution of probabilities in (3.2) as we did: *We wanted the specification of* $\mathfrak{p}_\omega$, $\omega \in \Omega$ *to make the coördinate random variables* X_1, X_2, $\cdots$, X_n *independent, under the resulting probability measure* $\mathbb{P}$ *of (3.3).* And this we have achieved.

In particular, with $x = (0, \cdots, 0, 1)$ we get from (3.5):

$$\mathbb{P}\big(X_1 = 0, \cdots, X_{n-1} = 0,\, X_n = 1\big) \,=\, p\,(1-p)^{\,n-1}\,. \tag{3.6}$$

- Finally, in a similar manner we obtain from (3.2) the **Binomial distribution**

$$\boxed{\mathbb{P}(S_n = k) \,=\, \sum_{\omega \in \Omega\,:\, \omega_1 + \cdots + \omega_n = k} \mathfrak{p}_\omega \,=\, \binom{n}{k}\, p^k\, (1-p)^{n-k}\,, \qquad k = 0, 1, \cdots, n} \tag{3.7}$$

for the total number of heads in n independent tosses of a coin, to wit, $S_n \sim \mathrm{Bin}\,(n;p)$. The binomial theorem guarantees $\sum_{k=1}^{n} \mathbb{P}(S_n = k) = 1$. It is checked readily from (3.7), using the binomial theorem once again, that $\mathbb{E}(S_n) = \sum_{k=0}^{n} k\,\mathbb{P}(S_n = k) = np$.

• We have constructed a probability space, and on it an entire *sequence* of independent random variables $X_1, X_2, \cdots$, with common BERNOULLI distribution as in (3.4). If we denote by $S_n = \sum_{i=1}^{n} X_i$ the sum of the first n of these random variables (that is, the total number of "heads" in the first n tosses of the coin) and by

$$\overline{X}_n := \frac{S_n}{n} \tag{3.8}$$

the relative frequency of heads in the first n tosses of the coin,

$$\lim_{n\to\infty} \mathbb{P}\left(|\,\overline{X}_n - p| \geq \varepsilon\right) = 0$$

holds for every $\varepsilon > 0$ (the "weak law of large numbers" of J. BERNOULLI (1713)), as does the demonstrably stronger statement

$$\mathbb{P}\left(\lim_{n\to\infty} \overline{X}_n = p\right) = 1$$

(the "strong law of large numbers" of CANTELLI (1917)). [18]

Exercise 3.3. *In a sequence of independent coin tosses, what is the probability that n successes (heads) materialize before m failures (tails) do?*

This question was posed around 1654 by an accomplished gambler, the Chevalier DE MÉRÉ, to the French philosopher, theologian and mathematician Blaise PASCAL.

Exercise 3.4. St. BANACH's Matchbox Problem: A smoker carries two matchboxes, one in the left pocket of his coat, the other in the right. Each time he lights his pipe, he chooses one pocket "at random" (meaning, with equal probability) and takes a match from the box in it. Initially, both boxes contain the same number n of matches. Eventually he finds that there are no matches left in his left pocket; what is the probability that there are k matches in the box in his right pocket?

3.3 POISSON Distribution

Let us now imagine a situation whereby, as the number of times n we toss a coin (cf. section 3.2) increases, the probability p of getting heads in a single toss decreases, say inversely proportionally to the length of the run: $p = \lambda/n$ for some constant $\lambda > 0$.

Making this substitution in the formula (3.7) for the binomial probabilities and letting $n \to \infty$, we obtain

$$\mathbb{P}(S_n = k) = \frac{n(n-1)\cdots(n-k+1)}{n^k}\,\frac{\lambda^k}{k!}\left(1-\frac{\lambda}{n}\right)^n\left(1-\frac{\lambda}{n}\right)^{-k} \longrightarrow e^{-\lambda}\,\frac{\lambda^k}{k!}$$

[18] The strong law of large numbers was first formulated by CANTELLI (1917); BOREL and HAUSDORFF had already discussed certain special cases. Like the weak law, this is only a very special case of general theorems on random variables, developed by KOLMOGOROV (1930) and BIRKHOFF (1932).

for $k = 0, 1, \cdots$. A random variable S with nonnegative integer values and

$$\mathbb{P}(S = k) \;=\; e^{-\lambda}\,\frac{\lambda^k}{k!} \;=:\; \mathfrak{p}_k\,, \qquad k = 0, 1, \cdots \tag{3.9}$$

is said to have the POISSON *(1837) distribution* with parameter $\lambda > 0$.

Here is a generalization of the considerations above.

Exercise 3.5. Show that we have convergence to the POISSON probabilities

$$\mathbb{P}(S_n = k) \;=\; \binom{n}{k}\,p_n^k\,\big(1 - p_n\big)^{n-k} \;\longrightarrow\; e^{-\lambda}\,\frac{\lambda^k}{k!} \;=\; \mathfrak{p}_k\;, \quad k = 0, 1, \cdots$$

also in the more general case $\lim_{n\to\infty}(n\,p_n) = \lambda > 0$.

Exercise 3.1 provides another concrete situation where this distribution arises, in fact with $\lambda = 1$ in that case. Indeed, the POISSON distribution comes up quite often, when both *large populations* and *unlikely, or "rare", events* are involved (the number of typos in the first draft of a long manuscript; the number of telephone calls passing through a given trunk in a given time-period; the number of times a given internet server is accessed during a given time-interval; the extremely rare coincidences of Exercise 3.1; et cetera).

3.4 Geometric Distribution

Imagine now that we start tossing a coin *as infinitum,* i.e., we take as our sample space the space $\Omega = \{0,1\}^{\mathbb{N}}$ which consists of all sequences $\omega = (\omega_1, \omega_2, \cdots)$ with all $\omega_i \in \{0,1\}$. We endow this space with an appropriate $\sigma-$algebra; [19] and on this $\sigma-$algebra we posit the existence of a probability measure $\mathbb{P}$ that satisfies the property (3.5) for any given $n \in \mathbb{N}$ and vector $x = (x_1, \cdots, x_n) \in \{0,1\}^n$.

It is not at all clear that such a measure should exist; this requires proof, and we shall return to this issue in due course. (A construction for the case $p = 1/2$ appears in Example 7.6.) Bravely assuming that we can do all of this, we consider the random variable $T : \Omega \to \mathbb{N} \cup \{\infty\}$ given by

$$T(\omega) \;:=\; \inf\big\{\, n \in \mathbb{N} \,:\, X_n(\omega) = 1 \,\big\}$$

if the indicated set is not empty, and by $T(\omega) := \infty$ if it is. *What is the distribution of this waiting time, the number of trials required until we see the first success?*

To answer this question, let us observe from (3.6) that for any $k \in \mathbb{N}$ we have

$$\mathbb{P}(T = k) \;=\; \mathbb{P}\big(X_1 = 0, \cdots, X_{k-1} = 0,\, X_k = 1\big) \;=\; p\,(1-p)^{k-1}\,.$$

This is the so-called *Geometric Distribution;* it satisfies

$$\mathbb{P}(T > k) \;=\; \mathbb{P}(T \geq k+1) \;=\; \mathbb{P}\big(X_1 = 0, \cdots, X_k = 0\big) \;=\; (1-p)^k$$

[19] E.g., the $\sigma-$algebra generated by the finite-dimensional cylinder sets, those that involve restrictions on a finite number of coördinates.

as well as the so-called *memoryless property*

$$\mathbb{P}\big(T > k+\ell \,|\, T > \ell\big) \;=\; \frac{\mathbb{P}\big(T > k+\ell\big)}{\mathbb{P}\big(T > \ell\big)} \;=\; (1-p)^k \;=\; \mathbb{P}(T > k)\,, \qquad \forall\ (k,\ell) \in \mathbb{N}^2\,. \tag{3.10}$$

It is seen readily that $\mathbb{P}(T=\infty) = 1 - \sum_{k\in\mathbb{N}} \mathbb{P}(T=k) = 0$. That is, although we allowed for the possibility that T might take the value ∞, this event $\{T=\infty\}$ has zero probability. We say that T is "almost surely" finite.

We shall make a habit of this: whenever a statement can fail only on a (subset of a) set that has $\mathbb{P}-$measure zero, we say that the statement holds *" $\mathbb{P}-$a.s. (almost surely)",* or for *" $\mathbb{P}-$a.e. (almost every) $\omega \in \Omega$."*

Let us also compute the expected time until the first success: from Exercise 2.7, we get

$$\mathbb{E}(T) \;=\; \sum_{k\in\mathbb{N}} \mathbb{P}(T \geq k) \;=\; \sum_{k\in\mathbb{N}} (1-p)^{k-1} \;=\; \frac{1}{1-(1-p)} \;=\; \frac{1}{p}\,.$$

• **A Doubling Strategy:** Now let us suppose that, while watching the successive outcomes of the coin toss, we are not above placing bets on the outcome of the next toss. Suppose indeed that we follow the "double-or-nothing" strategy of doubling our stakes after each "failure" (tails), until we see the first "success" (heads), and then calling it quits.

Under such a strategy, our fortune after the n^{th} toss is $M_n(\omega) = 0$, for $n = 0$; it is $M_n(\omega) = -\sum_{k=1}^{n} 2^{k-1} \;=\; 1 - 2^n$, for $1 \leq n < T(\omega)$; and $M_n(\omega) \;=\; 2\cdot 2^{n-1} + (1-2^n) \;=\; 1$, for $n \geq T(\omega)$. In other words, for $n \geq 1$ the random variable M_n takes only two values: the value 1 with probability $\mathbb{P}(T \leq n) = 1 - (1-p)^n$, and the value $1-2^n$ with probability $\mathbb{P}(T > n) = (1-p)^n$. Thus, for $n \geq 1$ the expectation of the random variable M_n is

$$\mathbb{E}(M_n) \;=\; 1 - (1-p)^n + (1-2^n)\cdot(1-p)^n \;=\; 1 - \big(2\,(1-p)\big)^n\,.$$

If the coin is fair, that is, if $p = 1/2$, this expectation is equal to zero, reflecting the fact that such a strategy is "neutral on the average" (you cannot win over a finite time horizon by betting, without clairvoyance of future events, on the outcome of a random sequence which has no clear tendency to go up or down). However, for a.e. $\omega \in \Omega$ (that is, for every ω in a set A with $\mathbb{P}(A) = 1$) we have

$$M_n(\omega) = 1 \;\text{ for all } n \in \mathbb{N} \text{ sufficiently large, simply because } \mathbb{P}(T < \infty) = 1\,.$$

To put all this into English: such a strategy will ultimately lead you to a gain. But because the sequence $\big\{M_n(\omega)\big\}_{n\in\mathbb{N}}$ is unbounded from below, you may have to absorb and sustain unbearable losses in the meantime; that is, you need the backing of a creditor with very deep pockets!

To quantify these statements, observe that with $p = 1/2$ we have $\mathbb{E}(T) = \sum_{n\in\mathbb{N}} n\,2^{-n} = 2$, so the wait is not exactly exorbitant; but the capital, or collateral, needed to back this strategy up, is $K = 2^{T-1} - 1$ and has expected value

$$\mathbb{E}(K) \;=\; \sum_{n\in\mathbb{N}} \big(2^{n-1}-1\big)\,\mathbb{P}(T=n) \;=\; \sum_{n\in\mathbb{N}} \big(2^{n-1}-1\big)\,2^{-n} \;=\; +\infty\,.$$

3.5 Gaussian Distribution

Let us recall the coin-tossing Example of section 3.2, along with the Binomial probabilities of (3.7), namely $\mathbb{P}(S_n = k) = \big(n!/k!(n-k)!\big)\, p^k\,(1-p)^{n-k}$. It can be shown that

$$\mathbb{P}\left[\, a \,\leq\, \frac{S_n - np}{\sqrt{np(1-p)}} \,\leq\, b \,\right] \,=\, \sum_{k \in \mathcal{K}_n} \mathbb{P}(S_n = k) \,\longrightarrow\, \Phi(b) - \Phi(a)\,, \qquad a < b \;\; \text{in } \mathbb{R} \tag{3.11}$$

as $n \to \infty$, where

$$\mathcal{K}_n \,:=\, \left\{ 1 \leq k \leq n \,:\, np + a\sqrt{np(1-p)} \,\leq k \leq np + b\sqrt{np(1-p)} \,\right\}$$

and

$$\Phi(x) \,:=\, \int_{-\infty}^{x} \varphi(\xi)\, \mathrm{d}\xi\,, \qquad \varphi(x) \,:=\, \frac{e^{-x^2/2}}{\sqrt{2\pi}} \tag{3.12}$$

is the so-called *standard Normal distribution function,* usually abbreviated $\mathcal{N}(0,1)$.

This remarkable fact was observed already in the 18^{th} century, first by Abraham DE MOIVRE (1730) in the case $p = 1/2$, and then by Pierre Simon, Marquis DE LAPLACE (1812) in the general case. It is a very special case of a far more general result of central importance in the theory and practice of Probability, the "Central Limit Theorem" of chapter 11. [20]

A random variable Z with distribution

$$\mathbb{P}(Z \in B) \,=\, \int_B \frac{e^{-x^2/2}}{\sqrt{2\pi}}\, dx \,=\, \int_B \varphi(x)\, dx\,, \qquad B \in \mathcal{B}(\mathbb{R})$$

is called *standard Gaussian* (or "normal"). More generally, a random variable $\widetilde{Z}$ that satisfies

$$\mathbb{P}(\widetilde{Z} \in B) \,=\, \int_B \frac{e^{-(x-m)^2/2\sigma^2}}{\sqrt{2\pi\sigma^2}}\, dx \,=\, \int_B \frac{1}{\sigma}\, \varphi\left(\frac{x-m}{\sigma}\right) dx\,, \qquad B \in \mathcal{B}(\mathbb{R})$$

for some constants $m \in \mathbb{R}$ and $\sigma > 0$, is said to have *Gaussian distribution* $\mathcal{N}(m, \sigma^2)$ centered at m and with dispersion parameter $\sigma > 0$. Such a random variable obviously has a diffuse distribution, with probability density function $(1/\sigma)\, \varphi((x-m)/\sigma)$ and probability distribution function $\Phi((x-m)/\sigma)$, $x \in \mathbb{R}$.

---

[20] It took two centuries, from the initial observation of DE MOIVRE (1730), until the general statement of what we call now "Central Limit Theorem" was proved by LINDEBERG (1922) and FELLER (1935). In the course of these centuries, successively more general versions of the result were being proved – until the full "universality" of the phenomenon was understood. In its most concrete version, this general result states that the convergence

$$\mathbb{P}\left[\, a \,\leq\, \frac{S_n - nm}{\sqrt{n\,\sigma^2}} \,\leq\, b \,\right] \longrightarrow\, \Phi(b) - \Phi(a)\,, \qquad a < b \;\; \text{in } \mathbb{R}$$

holds with $S_n = \sum_{j=1}^{n} X_j$ and a sequence $X_1, X_2, \cdots$ of independent copies of some random variable X with *arbitrary* distribution – as long as this distribution has finite variance $\sigma^2 \in (0, \infty)$ and expectation m. Only these two features of the distribution appear in the statement of the result – nothing else.

For a fascinating account of the history of this two-century struggle to understand this phenomenon, and the amazing list of great names involved in it, please consult the book by FISCHER (2011).

Remark 3.1. From the tables of the standard Gaussian distribution, one sees $\mathbb{P}(Z \in [-3,3]) = 2\Phi(3) - 1 \cong 99.997\%$. Thus, in the notation of (3.8), the event

$$\left\{ \frac{\overline{X}_n - p}{b(n)} \le K \right\}, \qquad b(n) := \sqrt{\frac{p(1-p)}{n}}$$

has probability strictly less than one for every $K \in (0,\infty)$; however, from (3.11) this probability is extremely close to one for $K = 3$ and n large. In other words: the convergence of the "relative frequency" $\overline{X}_n$ to the "true frequency" p is of the order $\sqrt{1/n}$, i.e., slow.

Note also that, for every $\varepsilon > 0$, the result (3.11) makes plausible the approximation

$$\mathbb{P}\left(|\,\overline{X}_n - p| \le \varepsilon\right) = \mathbb{P}\left(|\,S_n - np| \le \varepsilon n\right) \sim \int_{-\varepsilon/b(n)}^{\varepsilon/b(n)} \varphi(\xi)\, d\xi = 2\,\Phi\big(\varepsilon/b(n)\big) - 1 \longrightarrow 1 \quad (3.13)$$

as $n \to \infty$ in the notation of (3.8), "justifying" a claim made in section 3.2. Here and below, the sign "$\sim$" indicates that the ration of the quantities tends to one, as $n \to \infty$. In fact, we have now all the necessary ingredients required to turn these heuristics into a rigorous argument.

Exercise 3.6. Justify the heuristics of (3.13).

Exercise 3.7. Show the validity of the celebrated STIRLING *formula*

$$\boxed{n! \sim \sqrt{2\pi n}\; n^n\, e^{-n}}$$

and its stronger form

$$n\,! = \sqrt{2\pi}\; n^{n+1/2}\, e^{-n+\delta_n} \qquad \text{with} \qquad \frac{1}{12n+1} < \delta_n < \frac{1}{12n}\,. \quad (3.14)$$

Exercise 3.8. Establish the claim (3.11) made above.
(*Hint:* With the help of the STIRLING formula (3.14), show first the "local" form

$$\lim_{n\to\infty} \left(\sqrt{np(1-p)} \cdot \binom{n}{k_n(x)}\, p^{\,k_n(x)} (1-p)^{\,n-k_n(x)} \right) = \varphi(x) = \frac{e^{-x^2/2}}{\sqrt{2\pi}}$$

of this result, where $x \in \mathbb{R}$ is fixed and

$$k_n(x) := x\sqrt{np(1-p)} + np\,.$$

Then observe that this convergence is uniform over x in the bounded interval $[a,b]$ as the next step towards (3.11).)

3.6 Multinomial Distribution

Let us generalize the setting of Example 3.2: instead of tossing a coin, imagine casting a polyhedron with d facets $\mathfrak{f}_1, \cdots, \mathfrak{f}_d$ (say, a die if $k = 6$) a specified number n of times. Thus the sample space $\Omega = \{\mathfrak{f}_1, \cdots, \mathfrak{f}_d\}^n$ has cardinality $|\Omega| = d^n$, and for a typical element $\omega = (\omega_1, \cdots, \omega_n)$ we denote by $X_t(\omega) := \omega_t$ the outcome of the t^{th} toss: $\omega_t = \mathfrak{f}_i$ if the i^{th} facet turns up on the t^{th} toss. Similarly, we denote by

$$S_n^{(i)}(\omega) \; := \; \sum_{t=1}^{n} \mathbf{1}_{\{\mathfrak{f}_i\}}(\omega_t)$$

the number of occurrences of the i^{th} facet in n tosses of the polyhedron.

We specify a probability space by taking $\mathcal{F}$ to consist of all subsets of Ω, and by setting $\mathbb{P}(A) := \sum_{\omega \in A} \mathfrak{p}_\omega$ for any $A \subset \Omega$ where

$$\mathfrak{p}_\omega \; := \; p_1^{S_n^{(1)}(\omega)} \cdots p_d^{S_n^{(d)}(\omega)} \; , \qquad \omega \in \Omega$$

for some given numbers $p_1 > 0, \cdots, p_d > 0$ with $p_1 + \cdots + p_d = 1$. Intuitively, p_i is the probability that the facet $\mathfrak{f}_i$ turns up on any given toss; for instance, $p_1 = \cdots = p_6 = 1/6$ if we toss a die with all its facets equally likely.

For fixed integers $r_1 \geq 0, \cdots, r_d \geq 0$ with $r_1 + \cdots + r_d = n$, we have then

$$\mathbb{P}\left(S_n^{(1)} = r_1, \cdots, S_n^{(d)} = r_d\right) \; = \; \frac{n!}{r_1!\, r_2! \cdots r_d!} \cdot p_1^{r_1} \cdots p_d^{r_d} \, . \tag{3.15}$$

We say that the random vector $\left(S_n^{(1)}, \cdots, S_n^{(d)}\right)$ has a *Multinomial Distribution*.

Here in (3.15) the multinomial coëfficient counts the number of outcomes in Ω that result in r_1 appearances of facet $\mathfrak{f}_1$, in r_2 appearances of facet $\mathfrak{f}_2$, and so on to r_d appearances of facet $\mathfrak{f}_d$ (cf. section 13.6). It can checked as in section 3.2 that the random variables $X_1, \cdots, X_n$ are independent with common distribution $\mathbb{P}(X_t = \mathfrak{f}_i) = p_i$ for all $t = 1, \cdots, n$ and $i = 1, \cdots, d$.

It is also checked fairly easily that for $1 \leq i, j \leq n$ with $i \neq j$, the random variable $S_n^{(i)}$ has the Binomial distribution

$$\mathbb{P}\left(S_n^{(i)} = r_i\right) \; = \; \binom{n}{r_i} \cdot p_i^{r_i} \left(1 - p_i\right)^{n - r_i} , \qquad r_i = 0, 1, \cdots, n$$

whereas the random vector $\left(S_n^{(i)}, S_n^{(j)}\right)$ has the Trinomial distribution

$$\mathbb{P}\left(S_n^{(i)} = r_i \, , \; S_n^{(j)} = r_j\right) \; = \; \frac{n!}{r_i!\, r_j!\, (n - r_i - r_j)!} \cdot p_i^{r_i}\, p_j^{r_j} \left(1 - p_i - p_j\right)^{n - r_i - r_j}$$

for $r_i \geq 0$, $r_j \geq 0$ with $r_i + r_j \leq n$. By the same token, it is seen that any subcollection of $S_n^{(1)}, \cdots, S_n^{(d)}$ has a distribution of the Multinomial type.

3.7 Hypergeometric Distribution

An urn contains $n = r + b$ balls, out of which $r \geq 1$ are red and $b \geq 1$ are black. We are blindfolded, are asked to select m $(1 \leq m \leq n)$ balls "at random", meaning that all possible outcomes are equally likely, and we record the number X of red bals in our sample. *What is the probability distribution of this random variable?*

Clearly, X can take values in $\{0, 1, \cdots, r \wedge m\}$. If k red balls are selected from the available r, the remaining $m - k$ balls have to be selected from the available b black. By the basic combinatorial principle, both these selections – the conjunction of which amounts to a realization of the event $\{X = k\}$ – can be made in

$$\binom{r}{k} \cdot \binom{b}{m-k}$$

ways. The total number of ways to select m balls from the n available in the urn, is $(n)_m/m!$. Therefore,

$$\mathbb{P}(X = k) \equiv \mathfrak{p}_k := \frac{\binom{r}{k}\binom{b}{m-k}}{\binom{r+b}{m}}, \qquad k = 0, 1, \cdots, r \wedge m.$$

The resulting distribution is called *Hypergeometric.*

3.8 Exponential Distribution

You enter your friend's office, and observe she is on the phone. What is your assessment of how long it will take her to hang up? Would it help you make this assessment if I were to tell you that she has been on the phone already for ten minutes? would you change this assessment, were I to tell you that she has in fact been on the phone for an hour?

Suppose we can model the duration of your friend's phone conversation as a random variable $T : \Omega \to (0, \infty)$ with a continuous distribution function $F(x) = \mathbb{P}(T \leq x)$, $x \in \mathbb{R}$ and $F(0) = 0$, $F(\infty) = 1$, on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$. It is then not such a bad assumption, to postulate the *memoryless property*

$$\mathbb{P}(T > t + s \,|\, T > s) = \mathbb{P}(T > t), \qquad \forall\ t \geq 0,\ s \geq 0$$

(cf. (3.10): knowing how long the conversation has been going on reveals nothing about the likelihood it might be over during the next minute-and-a-half).

With the notation $G(x) := 1 - F(x) = \mathbb{P}(T > x)$ this equation becomes

$$G(t + s) = G(t)\, G(s), \qquad \forall\ t \geq 0,\ s \geq 0.$$

The only continuous solutions $G : [0, \infty) \to [0, 1]$ to this equation, are of the form $G(t) = e^{-\lambda t}$ for some constant $\lambda > 0$. The resulting probability distribution function

$$F(t) = \mathbb{P}(T \leq t) = 1 - e^{-\lambda t} = \int_0^t \lambda\, e^{-\lambda s}\, \mathrm{d}s, \qquad 0 \leq t < \infty \tag{3.16}$$

is called the *Exponential Distribution* function with parameter $\lambda > 0$. The random variable T has then a diffuse distribution, with probability density function $f(t) = \lambda\, e^{-\lambda t}\, \mathbf{1}_{(0,\infty)}(t)\,,\ t \in \mathbb{R}\,.$

The figure on the following page displays the relative frequencies of waiting times of a large number of patients for a medical examination at a hospital, over a 6-month period. The fit to exponentiality is rather striking. [21]

• **Gamma Distribution:** Consider independent random variables $T_1, T_2, \cdots$ with common exponential distribution as in (3.16) for some given $\lambda > 0$. Then the sum $S_n = \sum_{j=1}^{n} T_j$ of these variables has the *Gamma* $\Gamma(\lambda, n)$ *distribution,* given via

$$\mathbb{P}\big(S_n > t\big) \;=\; \int_t^\infty \lambda\, e^{-\lambda s}\, \frac{(\lambda s)^{n-1}}{(n-1)!}\, \mathrm{d}s\,, \qquad t \geq 0\,. \tag{3.17}$$

For another representation of this distribution, see Exercise 3.10 below.

3.9 POISSON Process

Suppose we run a store. We open at time $t = 0$ (say, 9:00AM sharp) and wait for customers to come. The first customer arrives and time T_1, the second at time $S_2 := T_1 + T_2$, the third at time $S_3 := T_1 + T_2 + T_3$, and so on. We cannot predict these times in advance, of course, so we model them as random variables. [22]

Consider, in other words, a sequence $T_1, T_2, \cdots$ of random variables $T_j : \Omega \to (0, \infty)$ that we are prepared to think of a "inter-arrival times"; as well as a generated sequence

$$S_0 := 0\,, \qquad S_n := \sum_{j=1}^{n} T_j\,, \quad n \in \mathbb{N}$$

that we then interpret as the *actual arrival times* of customers (particles, cars) to our facility. At any time $t \in [0, \infty)$ we keep track, then, of the number of customers (particles, cars)

$$N(t) \;:=\; \max\big\{n \in \mathbb{N}_0 \,\big|\, S_n \leq t\big\}\,, \qquad 0 \leq t < \infty \tag{3.18}$$

that have already arrived (been recorded, been counted).

This is a prototypical example of what we call *Stochastic Process:* a family of random variables indexed by the time parameter $t \in [0, \infty)$. In this case each realization of this process has the form of a random staircase: a piecewise constant, right-continuous function, that increases by jumps of size 1 each time a new customer arrives (or a new particle is recorded). We call it a *Counting Process,* to register this fact.

• Consider now the very special, but also very important, case where the random variables $T_1, T_2, \cdots$ are *independent,* with common *exponential* distribution as in (3.16) with given $\lambda > 0$.

[21] I am grateful to my colleague Professor Avi MANDELBAUM at the Technion in Haifa, Israel, for giving me permission to use this picture.

[22] Instead of customers at a store, imagine having a GEIGER-MÜLLER counter that records emissions of radioactive particles; or a surveillance camera that records the passage of cars at an intersection.

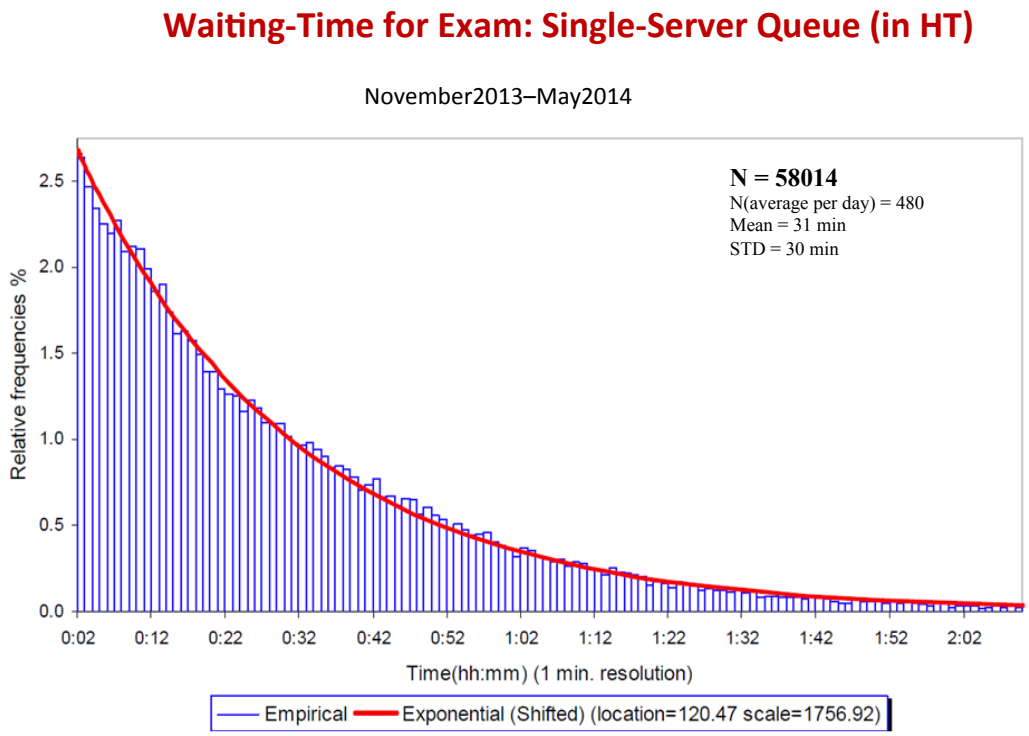

Figure 2: Distribution of Waiting Times for a Medical Examination at a Hospital. Courtesy of Professor Avi MANDELBAUM, Technion, Haifa.

The resulting counting process N of (3.18) is then called POISSON *Process* with parameter $\lambda > 0$, and has some rather remarkable properties: *It has stationary and independent increments, each of which has a* POISSON *distribution with parameter* $\lambda\, h$, where h is the length of the time-increment in question.

More precisely: suppose we fix time instances $t_0 = 0 \leq t_1 < t_2 < \cdots < t_m < \infty$. For the POISSON Process N, the random variables $N(t_1) - N(t_0)\,, \cdots\,, N(t_m) - N(t_{m-1})$ are then independent, and each $N(t_j) - N(t_{j-1})\,,\ j = 1, \cdots, m$ has POISSON distribution as in (3.9) with parameter $\lambda\,(t_j - t_{j-1})$. [23]

As a result, for any fixed real number $s \geq 0$, the random process $\big(N(s+t) - N(s)\big)_{0 \leq t < \infty}$ is *itself* POISSON with parameter $\lambda > 0$, and independent of $\big(N(u)\big)_{0 \leq u \leq s}$. This is the so-called MARKOV property of the POISSON process.

And a bit more generally: with S denoting any one of the "arrival times" $(S_n)_{n \in \mathbb{N}_0}$, the random process $\big(N(S+t) - N(S)\big)_{0 \leq t < \infty}$ is *itself* POISSON with parameter $\lambda > 0$, and independent of $\big(N(u)\big)_{0 \leq u \leq S}$. This is a foretaste of the so-called *Strong* MARKOV *property.*

Exercise 3.9. If N is a POISSON process with parameter $\lambda > 0$, and if $0 \leq s < t < \infty$ are given real numbers, compute

$$\mathbb{P}\big(N(s) = 1\,,\ N(t) = 2\big)\,.$$

Exercise 3.10. Show that, if N is a POISSON process with parameter $\lambda > 0$, and if $t > 0$ is a given real number, we have

$$\mathbb{P}\big(N(t) = n\big) \;=\; \mathbb{P}\big(S_n \leq t < S_{n+1}\big) \;=\; e^{\,-\lambda t}\,\frac{(\lambda t)^n}{n!}\,, \qquad k = 0, 1, \cdots\,.$$

(*Hint:* Argue by induction the validity of the representation

$$\mathbb{P}\big(S_n > t\big) \;=\; e^{\,-\lambda t} \sum_{k=0}^{n-1} \frac{(\lambda t)^k}{k!}\,, \qquad n = 1, 2, \cdots$$

for the Gamma $\Gamma(\lambda, n)$ distribution in (3.17).)

[23] This is quite tricky to prove rigorously; cf. BILLINGSLEY (1986), BRZEŹNIAK & ZASTAWNIAK (1999), pp. 145-148, and Problem 1.3.2 in KARATZAS & SHREVE (1991).

3.10 Uniform Distribution

A random variable with distribution function given as $F(x) = 0$ for $x \le a$, $F(x) = 1$ for $x \ge b$ and

$$F(x) \;=\; \frac{x-a}{b-a}\;, \qquad a < x < b$$

for some given $-\infty < a < b < \infty$, is said to have *uniform distribution* over the interval $[a, b]$. The distribution is diffuse, with probability density function $f(x) = (b-a)^{-1}\,\mathbf{1}_{(a,b)}(x)$.

Remark 3.2. *Preview of* VIETA: There is a neat way to obtain the uniform distribution over the interval $[-1, 1]$ starting from a sequence of BERNOULLI trials, as follows: Let $X_1, X_2, \cdots$ be independent random variables with common distribution $\mathbb{P}(X_k = 1) = p \in (0, 1)$, $\mathbb{P}(X_k = -1) = 1 - p$ for all $k \in \mathbb{N}$, and consider the sum

$$X \;=\; \sum_{n\in\mathbb{N}} X_n\, 2^{-n}\,.$$

- If $p = 1/2$, the random variable X has uniform distribution over $[-1, 1]$. [24]
- When $p \neq 1/2$, however, the distribution of the random variable X is continuous, strictly increasing on $[-1, 1]$ with $F(-1) = 0$, $F(1) = 1$, and differentiable with $F'(\cdot) \equiv 0$ away from a set of zero LEBESGUE measure. In other words, as soon as $p \neq 1/2$ with this simple structure, we have already an example of a continuous but non-diffuse distribution, and the inequality (2.5) becomes strict.

3.11 Tournaments

Consider a set $\mathcal{V}$ of n vertices, and on it a *Tournament* $\mathcal{T}_n$, that is, a complete directed graph. Equivalently, suppose we have n players, each of whom faces every other player, in a competition where no draws are allowed; we direct an edge from vertex (player) i to vertex (player) j, if i beats j. The schedule of the tournament does not matter, only the results. For a given integer $k < n$, we say that $\mathcal{T}_n$ *has property* $\mathcal{S}_k$, if for every set of k players $\{x_1, \cdots, x_k\}$, there is some other player y, who beats every player in the set.

SCHÜTTE was the first to pose the following problem: *Is it true that for every integer* k*, there exists a set* $\mathcal{V}$ *of* $n > k$ *vertices, and on it a Tournament* $\mathcal{T}_n$ *with the property* $\mathcal{S}_k$*?* And if so, what is the smallest necessary number $f(k)$ of players?

The answer of (3.19) below to SCHÜTTE's question was given by Paul ERDÖS (1963), and illustrates his "Probabilistic Approach" to Combinatorics and in Graph Theory. The rough idea is that, for $n \ge f(k)$ sufficiently large, an appropriately defined "random tournament" on the set $\mathcal{V} = \{1, \cdots, n\}$ of players is "very likely" to have the property $\mathcal{S}_k$.

For every $k \in \mathbb{N}$*, there exists a finite tournament* $\mathcal{T}_n$*,* $n > k$ *with the property* $\mathcal{S}_k$*.* (3.19)

[24] The easiest way to see this, is via characteristic functions; one obtains then the VIETA formula (7.15). Alternatively, establish (7.15) by trigonometric argumens, as we do in Example 7.6, then argue that X is uniformly distributed over $[-1, 1]$.

Proof of the claim (3.19): Consider a set $\mathcal{V} = \{1, \cdots, n\}$ of $n > k$ players, and a "random graph" $\mathcal{T}_n$ on it that corresponds to the idea of "deciding each game by tossing a coin independently from game to game". More formally, one takes as the sample space

$$\Omega = \big\{ (\omega_{ij})_{1 \le i,j \le n} \mid \omega_{ij} = 1 \ (i \text{ beats } j) \quad \text{or} \quad \omega_{ij} = -1 \ (j \text{ beats } i) \big\} ,$$

and assigns the probability 2^{-N}, $N = n(n-1)/2$ to each of its 2^N elements.

For any given subset $\mathcal{X} \subset \mathcal{V}$ with k elements, denote by $A_{\mathcal{X}}$ the property that "no player $y \in \mathcal{V} \setminus \mathcal{X}$ beats all players in $\mathcal{X}$". Clearly, $\mathbb{P}(v \text{ beats all players in } \mathcal{X}) = 2^{-k}$, for every $v \in \mathcal{V} \setminus \mathcal{X}$, and thus by independence:

$$\mathbb{P}(A_{\mathcal{X}}) = \big(1 - 2^{-k}\big)^{n-k} .$$

Therefore,

$$\mathbb{P}\Big(\bigcup_{\substack{\mathcal{X} \subset \mathcal{V} \\ \#\mathcal{X} = k}} A_{\mathcal{X}} \Big) \le \sum_{\substack{\mathcal{X} \subset \mathcal{V} \\ \#\mathcal{X} = k}} \mathbb{P}(A_{\mathcal{X}}) = \binom{n}{k} \big(1 - 2^{-k}\big)^{n-k} < 1 , \qquad \text{provided} \quad n \ge f(k) ,$$

where

$$f(k) := \min \left\{ m \ge k \;\middle|\; \binom{m}{k} \big(1 - 2^{-k}\big)^{m-k} < 1 \right\} .$$

In other words,

$$\mathbb{P}\Big(\bigcap_{\substack{\mathcal{X} \subset \mathcal{V} \\ \#\mathcal{X} = k}} (A_{\mathcal{X}})^c \Big) > 0 , \qquad \mathbb{P}\big(\text{the property } A_{\mathcal{X}} \text{ fails for all possible } \mathcal{X}\big) > 0 , \tag{3.20}$$

so *there is a point in the space* Ω (i.e., a tournament $\mathcal{T}_n$) with the property $\mathcal{S}_k$. □

It can be checked that $f(1) = 3$, $f(2) = 7$, and careful asymptotics give

$$c \cdot k 2^k \le f(k) \le k^2 2^k \, (1 + o(1)) , \qquad \text{as} \quad k \to \infty .$$

The argument leading to (3.20) is a typical illustration of P. ERDÖS's "probabilistic method": one shows the existence of an object possessing a certain property by constructing an appropriate probability space and showing that the event corresponding to the property under consideration has positive probability. The book by ALON & SPENCER (2000) is devoted to this subject.

3.12 The Exploits and Paradoxes of the Chevalier DE MÉRÉ

To close this chapter on a light note, let us recall the observation, listed in the Wednesday, July 29^{th}, 1654 letter from Blaise PASCAL to Pierre DE FERMAT – which is widely credited as having begotten the subject of Probability Theory:

It is more probable to get at least one six with four dice (let us denote this event by A), *than it is to get at least one double six in twenty four throws of two dice* (event B).

It is not hard for us today, to figure this out, assuming of course that repeated throws of dice are independent:

$$\mathbb{P}(A) = 1 - \left(\frac{5}{6}\right)^4 = 1 - \frac{625}{1296} = \frac{671}{625+671} = 0.5177\,, \qquad \mathbb{P}(B) = 1 - \left(\frac{35}{36}\right)^{24} = 0.4914\,.$$

Let us quote PASCAL himself: *"Mr.* DE MÉRÉ [25] *told me that he had found a fallacy in the theory of numbers, for this reason: if one undertakes to get a six with one die, the advantage of getting it in four throws is as 671 is to 625. If one undertakes to throw two sixes with four dice, there is a disadvantage in undertaking it in twenty-four throws".*

We have to grant it to the Chevalier DE MÉRÉ: He had figured out on his own *before* talking to PASCAL – by his smarts, powers of observation, or intuition (to say nothing of his long hours at the gambling table...) – the rather tight double inequality

$$\mathbb{P}(A) = 0.5177 > 0.50 > 0.4914 = \mathbb{P}(B)\,.$$

Now PASCAL continues: *"And nevertheless 24 is to 36 as 4 is to 6. This is what made him so indignant and made him say to one and all that the propositions are not consistent, and that Arithmetic was self-contradictory. But you will very easily see that what I say is correct, understanding the principles as you do".* [26] [27]

Exercise 3.11. Here is where the Chevalier probably went wrong: He apparently believed that "the laws of Arithmetic" mandated $\mathbb{P}(B)$ to be $24/36 = 0.6666$. Perhaps on the basis of a reasoning such as this: *"There are 36 possible outcomes when I throw two dice, one of which is favorable (the double-six), and I throw the two dice 24 times".*

Why is this reasoning spurious?

[25] The same fellow we encountered in Exercise 3.3.

[26] Please contrast the mental attitude towards randomness of Blaise PASCAL, a major mathematician but also great theologian and philosopher of the 17^{th} century, with the Late Antiquity mindset surrounding the story in Example 3.1. PASCAL does not seek explanations based on miracles, but rather on cold, rational assessment of the odds. His approach and attitude mark the dawn of the mindset that "made the world modern", as DEVLIN (2008) puts it.

[27] PASCAL and FERMAT also worked, in addition to the above "problem of dice", on the "problem of points", in which they had to wrestle with issues of what we would call today "fairness" in games of chance. Addressing this problem they truly went beyond what CARDANO and his great rival TARTAGLIA, or even GALILEO, had achieved, and arrived at the concept of what we call today "expectation".

4 The Basics of Measure and Integration

The first important property of the class of measurable functions is that it is *closed under limits*. To a very large extent it is this property that accounts for the significance of this concept, and for its centrality in the LEBESGUE theory of integration.

In fact, if $\{X_n\}_{n\in\mathbb{N}}$ is a sequence of real-valued functions, we can define the pointwise limit-superior and limit-inferior by

$$\limsup_{n\to\infty} X_n(\omega) := \inf_k \left(\sup_{n\geq k} X_n(\omega)\right), \quad \liminf_{n\to\infty} X_n(\omega) := \sup_k \left(\inf_{n\geq k} X_n(\omega)\right),$$

respectively. These quantities exist for every $\omega \in \Omega$, possibly with values $+\infty$ or $-\infty$, since the sequences $\left\{\sup_{n\geq k} X_n(\omega)\right\}_{k\in\mathbb{N}}$ and $\left\{\inf_{n\geq k} X_n(\omega)\right\}_{k\in\mathbb{N}}$ are monotone (decreasing and increasing, respectively).

For the purposes of this section, it is convenient then to consider functions $X : \Omega \to [-\infty, +\infty]$ with values in the extended real line. Such a function X is said to be *measurable*, if $X^{-1}((a, +\infty]) \in \mathcal{F}$ holds for every $a \in \mathbb{R}$ (equivalently, if $X^{-1}([-\infty, a)) \in \mathcal{F}$ holds for every $a \in \mathbb{R}$). Obviously, this definition agrees with the earlier one if X is finite everywhere.

Let us assume, therefore, that $\{X_n\}_{n\in\mathbb{N}}$ is a sequence of measurable functions on the measurable space $(\Omega, \mathcal{F})$, with values in the extended real line $[-\infty, +\infty]$. Then for any $k \in \mathbb{N}$ we have

$$\left(\sup_{n\geq k} X_n\right)^{-1}\big((a,\infty]\big) = \bigcup_{n\geq k} X_n^{-1}\big((a,\infty]\big) \quad [28]$$

$$\left(\inf_{n\geq k} X_n\right)^{-1}\big([-\infty,a)\big) = \bigcup_{n\geq k} X_n^{-1}\big([-\infty,a)\big),$$

so that the supremum and the infimum of a sequence of measurable functions are themselves measurable functions.

Iterating the argument, we deduce the following extremely important result.

Theorem 4.1. *The limit-inferior and the limit-superior of a sequence $\{X_n\}_{n\in\mathbb{N}}$ of measurable functions, are also measurable functions. In particular, if the sequence converges pointwise, its limit is a measurable function.*

Notation 4.1. The LEBESGUE Spaces: We shall denote throughout by $\mathbb{L}^0$ the space of measurable functions $X : \Omega \to \mathbb{R}$; by $\mathbb{L}^0_+$ the space of measurable functions $X : \Omega \to [0, \infty)$; and by $\mathbb{L}^*_+$ the space of measurable functions $X : \Omega \to [0, \infty]$.

Suppose that we endow now the measurable space $(\Omega, \mathcal{F})$ with **an arbitrary measure** $\mathbb{P}$, and define as in (2.6), (2.8) the integral $\mathbb{E}(X)$ of $X \in \mathbb{L}^*_+$. We denote by $\mathbb{L}^1$ the subspace of $\mathbb{L}^0$ that consists of all integrable functions $X : \Omega \to \mathbb{R}$, i.e., with $\mathbb{E}(|X|) < \infty$.

A bit more generally, for any given $p \in (0, \infty)$ we shall denote by $\mathbb{L}^p$ the subspace of $\mathbb{L}^0$ that consists of all measurable functions $X : \Omega \to \mathbb{R}$ with $\mathbb{E}\big(|X|^p\big) < \infty$.

[28] For the world record in pole vault ever to exceed 8m, someone has to clear that height, eventually.

Convention: Abusing notation a bit, we shall identify measurable functions that agree almost everywhere with respect to this measure $\mathbb{P}$; that is, $\mathbb{L}^0$ and its subspaces will be thought of as containing equivalence classes of measurable functions, as opposed to functions proper, where

$$X \sim Y \iff \mathbb{P}(X \neq Y) = 0\,.$$

For instance, on the probability space $(\Omega, \mathcal{F}, \mathbb{P}) = ([0,1], \mathcal{B}([0,1]), \lambda|_{[0,1]})$ we identify in the same equivalence class in $\mathbb{L}^0_+$ all the functions $X_k = k\,\mathbf{1}_{\mathbb{Q}\cap[0,1]}\,,\ k \in [0,\infty]$; in fact, we have $\mathbb{E}(X_k) = \mathbb{E}(X_0) = 0\,.$ [29]

We have the following three fundamental results, Theorems 4.2-4.4; they provide conditions under which we can interchange the operations of "limit" and "integral". Their generality and strength are direct consequences of the countable additivity of measure.

Theorem 4.2. B. LEVI's Monotone Convergence: *Suppose that the sequence $\{X_n\}_{n\in\mathbb{N}} \subset \mathbb{L}^*_+$ is monotone increasing, in the sense that $X_n \leq X_{n+1}$ holds pointwise for all $n \in \mathbb{N}$. Then*

$$\boxed{\lim_{n\to\infty} \mathbb{E}\big(X_n\big) \;=\; \mathbb{E}\big(\lim_{n\to\infty} X_n\big)\;.} \tag{4.1}$$

Note that neither side in (4.1) is assumed to be finite.

We cannot dispense in this result with the assumption of monotonicity. For instance, consider the functions $X_n(\omega) = n\,\mathbf{1}_{(0,1/n)}(\omega)\,,\ n \in \mathbb{N}$ on the probability space $\Omega = (0,1]$ with LEBESGUE measure on its BOREL sets. Observe that we have $\mathbb{E}(X_n) = 1$ for all $n \in \mathbb{N}$, that the sequence $\{X_n\}_{n\in\mathbb{N}}$ *is not monotone*, and that $X(\omega) := \lim_n X_n(\omega) \equiv 0$ "pointwise", that is, *for every* $\omega \in \Omega$: thus $\mathbb{E}(X) = 0\,.$ Note, however, that we have $\mathbb{E}(X) \leq \lim_n \mathbb{E}(X_n)$ in this example, echoing a more general result known as "FATOU's Lemma".

Theorem 4.3. FATOU's Lemma: *For any sequence $\{X_n\}_{n\in\mathbb{N}} \subset \mathbb{L}^*_+$ we have*

$$\boxed{\mathbb{E}\big(\liminf_{n\to\infty} X_n\big) \;\leq\; \liminf_{n\to\infty}\, \mathbb{E}(X_n)\,.}$$

Theorem 4.4. LEBESGUE's Dominated Convergence: *Let $\{X_n\}_{n\in\mathbb{N}} \subset \mathbb{L}^0$ be any sequence of measurable functions with $|X_n| \leq Y$ valid pointwise for all $n \in \mathbb{N}$, where $Y \in \mathbb{L}^1$ is an integrable measurable function.*

Then, if the sequence $\{X_n\}_{n\in\mathbb{N}}$ converges pointwise, the limit is integrable and (4.1) holds.

Rather nasty things can happen when the conditions of Theorems 4.2-4.4 fail, as the Exercise 4.18 illustrates. But when they are satisfied they can help establish beautiful results, as the following Exercises illustrate.

[29] Here is a big difference between the RIEMANN and the LEBESGUE theories of integration. In the RIEMANN theory the functions $X_k\,,\ k \in (0,\infty]$ are not integrable. In the LEBESGUE theory, not only are they integrable; in fact, they have the same integral as $X_0\,.$

Exercise 4.1. The Central Limit Theorem for the Heat Equation (COLDING (2020)): Imagine a rod of iron that extends to infinity, both to the left and to the right. At time $t = 0$, we specify an initial temperature profile: $f(y) \geq 0$ is the temperature at location $y \in \mathbb{R}$ on the iron rod.

Newton's Law of Cooling specifies that the temperature $u(t, x)$ at a subsequent time $t > 0$ and location $x \in \mathbb{R}$ has to satisfy the *Heat Equation*

$$\frac{\partial u}{\partial t} \;=\; c \cdot \frac{\partial^2 u}{\partial x^2}\,, \qquad (t, x) \in (0, \infty) \times \mathbb{R}$$

for some real constant $c > 0$ (depending on the specific heat and thermal conductivity of the material), and of course the initial condition $\lim_{t \downarrow 0} u(t, x) = f(x),\ x \in \mathbb{R}$.

Assume that the initial temperature profile $f : \mathbb{R} \to [0, \infty)$ is continuous and integrable with $0 < \int_{\mathbb{R}} f(y)\, \mathrm{d}y < \infty$, and take $c = 1/2$ for concreteness. Show that

$$u(t, x) = \int_{\mathbb{R}} p\big(t, x - y\big)\, f(y)\, \mathrm{d}y\,, \qquad \text{with} \qquad p(t, \xi) := \frac{e^{\,-\xi^2/(2t)}}{\sqrt{2\pi t}}$$

the probability density function of the GAUSSian distribution with mean zero and variance t, solves the heat equation and satisfies the initial condition. Show also that the resulting temperature profile $u(t, \cdot)$ at time $t > 0$ — appropriately scaled and normalized — approaches as $t \to \infty$ the standard GAUSSian probability density function; to wit,

$$\lim_{t \to \infty} \left(\sqrt{t}\ \frac{u\big(t, x\sqrt{t}\,\big)}{\int_{\mathbb{R}} f(y)\, \mathrm{d}y} \right) \;=\; \frac{e^{\,-x^2/2}}{\sqrt{2\pi}}\,, \qquad \forall\ x \in \mathbb{R}\,.$$

Exercise 4.2. SCHEFFÉ's Lemma: [30] On a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, consider a sequence of *integrable* random variables $\big(X_n\big)_{n \in \mathbb{N}}$ that converges a.e. to another *integrable* variable X.

(i) Suppose, in addition that all these random variables are non-negative: $X_n \geq 0$, $X \geq 0$. Show then, that

$$\lim_{n \to \infty} \mathbb{E}\big(|X_n|\big) \;=\; \mathbb{E}\big(|X|\big) \iff \lim_{n \to \infty} \mathbb{E}\big(|X_n - X|\big) \;=\; 0\,. \tag{4.2}$$

(*Hint:* You may find it useful to express the absolute value $|X - X_n|$ on the right-hand side of (4.2) in terms of its positive and negative parts. Of course, the absolute values $|X_n|$, $|X|$ are unnecessary on the left-hand side of (4.2) in this case of non-negative $X_n \geq 0$, $X \geq 0$.)

(ii) Now prove the equivalence in (4.2) in the general case, that is, *without* assuming non-negativity.

4.1 Composition and Change of Variable

We also have the following basic result. Though not exactly hard to prove, it gets used literally all the time when computing integrals and expectations.

[30] Henry SCHEFFÉ was one of the very first faculty members of the Mathematical Statistics Department here at Columbia, in the late 1940's.

Theorem 4.5. Composition and Change of Variable: *Let* $(\Omega, \mathcal{F})$ *and* $(\mathfrak{S}, \mathcal{G})$ *be two measurable spaces, suppose the mapping* $X : \Omega \to \mathfrak{S}$ *is measurable* (*that is,* $X^{-1}(G) \in \mathcal{F}$ *holds for every* $G \in \mathcal{G}$), *and for a given measure* $\mathbb{P}$ *on* $\mathcal{F}$ *define a measure* $\mu = \mathbb{P}\, X^{-1}$ *on* $\mathcal{G}$ *via*

$$\mu(G) \;:=\; \mathbb{P}\big(X^{-1}(G)\big)\,, \qquad G \in \mathcal{G}\,.$$

Consider also a real-valued, $\mathcal{G}-$*measurable function* $\mathfrak{h} : \mathfrak{S} \to \mathbb{R}$. *Then we have the following:*

(i) The composition $\mathfrak{h}\, X : \Omega \to \mathbb{R}$, *defined by* $(\mathfrak{h}X)(\omega) := \mathfrak{h}\big(X(\omega)\big)$, *is* $\mathcal{F}-$*measurable.*

(ii) If $\mathfrak{h}$ *is nonnegative, we have the* change-of-variable formula

$$\int_{X^{-1}(G)} \big(\mathfrak{h}\, X\big)(\omega)\, \mathrm{d}\mathbb{P}(\omega) \;=\; \int_G \mathfrak{h}(\xi)\, \mathrm{d}\mu(\xi)\,, \qquad \forall\; G \in \mathcal{G}\,. \tag{4.3}$$

In particular, $\mathbb{E}^{\mathbb{P}}(\mathfrak{h}\, X) = \int_{\mathfrak{S}} \mathfrak{h}\, \mathrm{d}\mu$.

(iii) More generally, $\mathfrak{h}$ *is integrable with respect to* μ *if and only if* $\mathfrak{h}X$ *is integrable with respect to* $\mathbb{P}$, *and under this condition the change-of-variable formula* (4.3) *is valid again.*

(iv) Consider now the special case where $(\Omega, \mathcal{F}, \mathbb{P})$ *is a* **probability space** *and* $(\mathfrak{S}, \mathcal{G}) = (\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$. *Consider also a real-valued,* $\mathcal{B}(\mathbb{R}^n))-$*measurable function* $\mathfrak{h} : \mathbb{R}^n \to \mathbb{R}$.

Then $\mu = \mathbb{P}\, X^{-1} \equiv \mu_X$ *is the distribution of the random vector* X, [31] *the composition* $\mathfrak{h}\, X : \Omega \to \mathbb{R}$ *is a random variable, and (4.3) becomes*

$$\boxed{\mathbb{E}[\,\mathfrak{h}(X)\,] \;=\; \int_\Omega \big(\mathfrak{h}\, X\big)(\omega)\, \mathrm{d}\mathbb{P}(\omega) \;=\; \int_{\mathbb{R}^n} \mathfrak{h}(\xi)\, \mathrm{d}\mu_X(\xi)\,.} \tag{4.4}$$

Exercise 4.3. Show that two random variables $X : \Omega \to \mathbb{R},\ Y : \Omega \to \mathbb{R}$, are identically distributed if, and only if, the identity

$$\mathbb{E}\big[\Psi(X)\big] \;=\; \mathbb{E}\big[\Psi(Y)\big]$$

of (2.16) holds for every bounded, continuous function $\Psi : \mathbb{R} \to \mathbb{R}$.

We express this colloquially by saying that the collection $\mathcal{C}_b(\mathbb{R})$ of bounded, continuous functions $\Psi : \mathbb{R} \to \mathbb{R}$ is a "distribution-determining class".

4.1.1 Moments and Generating Functions; Examples

The formula (4.4) is very important in Probability and Statistics: so much so that it is often mistaken as a definition (the so-called "law of the unconscious statistician"). It allows the almost mechanical calculation of the

- *moments:* $\mathbb{E}(X^k)$, $k = 1, 2, 3, \cdots$,
- *variances:*
$$\mathrm{Var}(X) := \mathbb{E}\big(X - (\mathbb{E}(X))\big)^2 = \big(\mathbb{E}(X^2)\big) - \big(\mathbb{E}(X)\big)^2\,,$$
- *moment generating functions:*
$$\phi_X(\xi) := \mathbb{E}\big(e^{\xi\, X}\big)\,,\ \ \xi \in \mathbb{R}\,,$$

[31] This is the induced measure of (2.3), when $n = 1$.

of a random variable X, whenever these quantities exist; as well as of the

- *covariances:*

$$\mathrm{Cov}(X_i, X_j) := \mathbb{E}\big[\big(X_i - \mathbb{E}(X_i)\big)\big(X_j - \mathbb{E}(X_j)\big)\big] = \mathbb{E}\big(X_i X_j\big) - \mathbb{E}(X_i)\,\mathbb{E}(X_j)\,,$$

- *moment generating functions:*

$$\phi_X(\theta) := \mathbb{E}\big(e^{\,\theta_1 X_1 + \cdots + \theta_n X_n}\big)\,,\ \ \theta = (\theta_1, \cdots, \theta_n) \in \mathbb{R}^n\,,$$

for a vector $X = (X_1, \cdots, X_n)$ of random variables.

We say that two random variables X, Y are *uncorrelated,* if $\mathrm{Cov}(X, Y) = 0$.

Exercise 4.4. Show that for the Binomial distribution of subsection 3.2 we have $\mathbb{E}(X) = np$, $\mathrm{Var}(X) = np(1-p)$, $\phi_X(\theta) = \big(1 - p + p\,e^{\,\theta}\big)^n$.

Exercise 4.5. Show that for the Poisson distribution of subsection 3.3 we have $\mathbb{E}(X) = \lambda$, $\mathrm{Var}(X) = \lambda$, $\phi_X(\theta) = \exp\big(\lambda(e^{\,\theta} - 1)\big)$.

Exercise 4.6. Show that for the Gaussian $\mathcal{N}(m, \sigma^2)$ distribution of subsection 3.5 we have $\mathbb{E}(X) = m$, $\mathrm{Var}(X) = \sigma^2$, $\phi_X(\theta) = \exp\big(m\,\theta + (\sigma^2/2)\,\theta^2\big)$.

Exercise 4.7. Show that for the Multinomial distribution of subsection 3.6 we have $\mathrm{Cov}(X_k, X_\ell) \,=\, -n\,p_k\,p_\ell$ for $1 \le k, \ell \le n$.

Exercise 4.8. Show that for the Exponential distribution of subsection 3.8 we have $\mathbb{E}(X) = 1/\lambda$, $\mathrm{Var}(X) = 1/\lambda^2$, $\mathbb{E}(X^k) = \Gamma(k+1)/\lambda^k$, $\phi_X(\theta) = \lambda/(\lambda - \theta)$ for $\theta < \lambda$.

Exercise 4.9. Consider random variables $X_1, \cdots, X_n$ in $\mathbb{L}^2$. Show that we have

$$\mathrm{Var}\Big(\sum_{j=1}^n X_j\Big) \,=\, \sum_{j=1}^n \mathrm{Var}(X_j) + 2\sum_{j=1}^n \sum_{i=j+1}^n \mathrm{Cov}(X_i, X_j)\,.$$

Observe also, that if the $X_j = \mathbf{1}_{A_j}$, $j = 1, \cdots, n$ are indicators, then

$$\mathrm{Var}\Big(\sum_{j=1}^n X_j\Big) \,\le\, \mathbb{E}\Big(\sum_{j=1}^n X_j\Big) + 2\sum_{j=1}^n \sum_{i=j+1}^n \mathrm{Cov}(X_i, X_j)\,.$$

Exercise 4.10. Compute the expectation and the variance of the Hypergeometric distribution in subsection 3.7.

Let us do some computations in the case of the POISSON distribution, by way of illustration. On the strength of the expression (4.4), we have

$$\mathbb{E}(X) \,=\, \sum_{k\in\mathbb{N}} k \cdot e^{-\lambda}\frac{\lambda^k}{k!} \,=\, \lambda\, e^{-\lambda} \sum_{k\in\mathbb{N}} \frac{\lambda^{k-1}}{(k-1)!} \,=\, \lambda$$

as well as $\mathbb{E}(X^2) = \lambda^2 + \lambda$ and thus $\mathrm{Var}(X) = \lambda$, since

$$\mathbb{E}\big(X(X-1)\big) \,=\, \sum_{k\ge 2} k(k-1) \cdot e^{-\lambda}\frac{\lambda^k}{k!} \,=\, \lambda^2\, e^{-\lambda} \sum_{k\ge 2} \frac{\lambda^{k-2}}{(k-2)!} \,=\, \lambda^2\,.$$

4.2 Limits Inferior and Superior for Sequences of Sets

Let us start by recalling, for an arbitrary sequence $\{E_n\}_{n\in\mathbb{N}}$ of measurable sets, the definitions of the *limit-inferior*

$$\boxed{\liminf_{n\to\infty} E_n \;\equiv\; \{E_n\,,\ \text{ev.}\} \;:=\; \bigcup_{k\in\mathbb{N}} \Big(\bigcap_{n\geq k} E_n \Big)\,,} \tag{4.5}$$

and the *limit-superior*

$$\boxed{\limsup_{n\to\infty} E_n \;\equiv\; \{E_n\,,\ \text{i.o.}\} \;:=\; \bigcap_{k\in\mathbb{N}} \Big(\bigcup_{n\geq k} E_n \Big) \;:=\; \Big\{ \sum_{n\in\mathbb{N}} \mathbf{1}_{E_n} = \infty \Big\}\,,} \tag{4.6}$$

By definition, an element $\omega\in\Omega$ belongs to the "smaller" set $\liminf_{n\to\infty} E_n$ of (4.5), iff $\omega\in E_n$ holds for all but finitely many integers $n\in\mathbb{N}$ or, as we say, *"eventually"* (abbreviated *"ev."*); equivalently, iff there exists an integer $K(\omega)\in\mathbb{N}$ such that $\omega\in E_n$ holds for all $n\geq K(\omega)$.

By definition, an element $\omega\in\Omega$ belongs to the "larger" set $\limsup_{n\to\infty} E_n$ of (4.6), iff $\omega\in E_n$ holds for infinitely many $n\in\mathbb{N}$ or, as we say, *"infinitely often"* (abbreviated *"i.o."*); equivalently, iff for every integer $k\in\mathbb{N}$ there exists an integer $n=N(\omega)\geq k$ such that $\omega\in E_n$.

We have clearly $\big(\liminf_{n\to\infty} E_n\big)^c = \limsup_{n\to\infty} E_n^c\,,$ as well as

$$\liminf_{n\to\infty} E_n \;\subseteq\; \limsup_{n\to\infty} E_n\,;$$

and when equality hods in the above display, we say that the sequence $\{E_n\}_{n\in\mathbb{N}}$ *converges*, and denote the common value $\lim_{n\to\infty} E_n$. This happens, for instance, when the sequence $\{E_n\}_{n\in\mathbb{N}}$ is monotone (increasing, or decreasing), as we now illustrate.

Remark 4.1. If the sequence $\{E_n\}_{n\in\mathbb{N}}$ is increasing ($E_n\subseteq E_{n+1}\,,\ \forall\, n\in\mathbb{N}$), we have $\bigcap_{n\geq k} E_n = E_k$ and $\bigcup_{n\geq k} E_n = \bigcup_{n\in\mathbb{N}} E_n$ for all $k\in\mathbb{N}$, thus

$$\liminf_{n\to\infty} E_n \;=\; \limsup_{n\to\infty} E_n \;=\; \bigcup_{n\in\mathbb{N}} E_n \equiv \lim_{n\to\infty} E_n\,.$$

Whereas, if the sequence $\{E_n\}_{n\in\mathbb{N}}$ is decreasing ($E_{n+1}\subseteq E_n\,,\ \forall\, n\in\mathbb{N}$), then we have $\bigcup_{n\geq k} E_n = E_k$ and $\bigcap_{n\geq k} E_n = \bigcup_{n\in\mathbb{N}} E_n$ for all $k\in\mathbb{N}$, thus

$$\liminf_{n\to\infty} E_n \;=\; \limsup_{n\to\infty} E_n \;=\; \bigcap_{n\in\mathbb{N}} E_n \;\equiv\; \lim_{n\to\infty} E_n\,.$$

Exercise 4.11. Limits Superior and Inferior; a foretaste of BOREL-CANTELLI: With these definitions, we have the following properties:

(i): $\mathbb{P}\left(\liminf_{n\to\infty} E_n\right) \;\leq\; \liminf_{n\to\infty} \mathbb{P}\left(E_n\right)$*;*

(ii): $\mathbb{P}\left(\limsup_{n\to\infty} E_n\right) \;\geq\; \limsup_{n\to\infty} \mathbb{P}\left(E_n\right)$, *if* $\mathbb{P}\left(\bigcup_{n\geq k} E_n\right) < \infty$ *for some* $k\in\mathbb{N}$*;*

(iii): $\mathbb{P}\left(\limsup_{n\to\infty} E_n\right) = 0$ *if* $\sum_{n\in\mathbb{N}} \mathbb{P}(E_n) < \infty$.

The property (iii) is known as the *First* BOREL-CANTELLI *Lemma.*

Exercise 4.12. Determine the sets of (4.5), (4.6) when the sequence of sets $\{E_n\}_{n\in\mathbb{N}}$ is specified as follows:

(i): $E_n = [a_n, b_n)$ where, for each $n \in \mathbb{N}$ we have $0 < a_n < b_n$, $b_n > 1$ and $\lim_{n\to\infty} a_n = 0$, $\lim_{n\to\infty} b_n = 1$.

(ii): $E_n = \{m/n \,:\, m \in \mathbb{N}\}$ for each $n \in \mathbb{N}$.

4.3 Proofs of Theorems 4.1-4.5

We begin by establishing Theorem 4.2. As a preparatory step, let us consider the case where each X_n is the indicator function of a measurable set E_n, namely, $X_n = \mathbf{1}_{E_n}$ for all $n \in \mathbb{N}$. Then Theorem 2.1 is equivalent to the following property of measure:

• **Continuity from Below:** *Whenever $\{E_n\}_{n\in\mathbb{N}} \subseteq \mathcal{F}$ is a monotone increasing sequence of measurable sets, i.e., $E_n \subseteq E_{n+1}$ holds for every $n \in \mathbb{N}$, we have*

$$\boxed{\lim_{n\to\infty} \mathbb{P}(E_n) = \mathbb{P}\Big(\bigcup_{n\in\mathbb{N}} E_n\Big)\,.} \tag{4.7}$$

To establish this property, imagine peeling an onion – that is, express each E_n as a finite disjoint union $E_n = \bigcup_{k=1}^{n} F_k$ with $F_1 = E_1$, and $F_k = E_k \setminus E_{k-1}$ for $k \geq 2$ (the successive layers of the onion); in particular, we have then $\bigcup_{n\in\mathbb{N}} E_n = \bigcup_{k\in\mathbb{N}} F_k$ and $\mathbb{P}(E_n) = \sum_{k=1}^{n} \mathbb{P}(F_k)$ for $n \in \mathbb{N}$. The countable additivity of the measure $\mathbb{P}$ implies

$$\mathbb{P}\Big(\bigcup_{n\in\mathbb{N}} E_n\Big) = \mathbb{P}\Big(\bigcup_{k\in\mathbb{N}} F_k\Big) = \sum_{k\in\mathbb{N}} \mathbb{P}(F_k) = \lim_{n\to\infty} \sum_{k=1}^{n} \mathbb{P}(F_k) = \lim_{n\to\infty} \mathbb{P}(E_n)\,.$$

• Returning to the proof of Theorem 4.2, let us recall the monotonicity property (2.9) of the integral $\mathbb{E}(X)$ for $X \in \mathbb{L}^0_+$, which gives

$$\int_\Omega X_n \,\mathrm{d}\mathbb{P} \,\leq\, \int_\Omega X_{n+1} \,\mathrm{d}\mathbb{P} \,\leq\, \int_\Omega X \,\mathrm{d}\mathbb{P}\,, \quad \forall\, n \in \mathbb{N}$$

since $X_n \leq X_{n+1} \leq \lim_{n\to\infty} X_n =: X$. The sequence $\big\{\int_\Omega X_n \, d\mathbb{P}\big\}_{n\in\mathbb{N}}$ is increasing, and hence has a limit in $[0, \infty]$. In view of this, we obtain

$$\lim_{n\to\infty} \int_\Omega X_n \,\mathrm{d}\mathbb{P} \,\leq\, \int_\Omega X \,\mathrm{d}\mathbb{P}\,.$$

To prove the reverse inequality, we fix momentarily a number $c \in (0, 1)$ as well as a simple function Y satisfying $0 \leq Y \leq X$, and set

$$E_n := \big\{\omega \in \Omega \,|\, X_n(\omega) \geq cY(\omega)\big\} \equiv \{X_n \geq cY\}.$$

Then we have $E_n \subseteq E_{n+1}$ and $\mathbf{1}_{E_n} \leq \mathbf{1}_{E_{n+1}}$ pointwise for all $n \in \mathbb{N}$, as well as $\Omega = \bigcup_{n\in\mathbb{N}} E_n$. The homogeneity and monotonicity properties (2.7), (2.9) of the integral imply

$$c \int_{E_n} Y \, d\mathbb{P} \,= c \cdot \mathbb{E}\,(Y\,\mathbf{1}_{E_n}) = \mathbb{E}\,(c \cdot Y\,\mathbf{1}_{E_n}) \equiv \int_{E_n} c\,Y \,\mathrm{d}\mathbb{P} \,\leq\, \int_{E_n} X_n \,\mathrm{d}\mathbb{P} \,\leq\, \int_\Omega X_n \,\mathrm{d}\mathbb{P}\,.$$

The limit as $n \to \infty$ of the sequence of integrals $\{\mathbb{E}(Y\,\mathbf{1}_{E_n})\}_{n\in\mathbb{N}}$ exists, and equals

$$\lim_{n\to\infty}\int_{E_n} Y\,\mathrm{d}\mathbb{P} = \lim_{n\to\infty}\sum_{k=1}^{K} y_k\cdot\mathbb{P}\left(Y^{-1}(\{y_k\})\cap E_n\right) = \sum_{k=1}^{K} y_k\cdot\mathbb{P}\left(Y^{-1}(\{y_k\})\right) = \int_\Omega Y\mathrm{d}\mathbb{P}$$

thanks to the continuity-from-below property of (2.5); here $\{y_1,\cdots,y_K\}$ is the range of the simple function Y. Thus we obtain

$$c\int_\Omega Y\,\mathrm{d}\mathbb{P} \;\le\; \lim_{n\to\infty}\int_\Omega X_n\,\mathrm{d}\mathbb{P}\,.$$

Taking the supremum over all simple functions Y satisfying $0 \le Y \le X$, and then over all $c\in(0,1)$, yields Theorem 4.2 on the strength of $\int_\Omega X\,d\mathbb{P} \le \lim_{n\to\infty}\int_\Omega X_n\,\mathrm{d}\mathbb{P}$. □

• Let us consider FATOU's lemma next. Recall that $\liminf_{n\to\infty}X_n$ is the pointwise limit of the monotone increasing sequence of functions $H_k := \inf_{n\ge k}X_n$, $k\in\mathbb{N}$. From $H_k \le X_k$, the monotonicity of the integral gives $\mathbb{E}(H_k)\le\mathbb{E}(X_k)$; from Theorem 4.2 we obtain then

$$\int_\Omega \left(\lim_{k\to\infty}H_k\right)\mathrm{d}\mathbb{P} \;=\; \lim_{k\to\infty}\int_\Omega H_k\,d\mathbb{P} \;\le\; \liminf_{k\to\infty}\int_\Omega X_k\,\mathrm{d}\mathbb{P}\,,$$

and FATOU's lemma is proved. □

• We turn now to the proof of Theorem 4.4. It is instructive to see first, why *the requirement that all the $|X_n|$ be dominated by a fixed integrable function Y cannot be removed.*

Consider again the case where each $X_n = \mathbf{1}_{E_n}$ is the indicator function of a set $E_n\in\mathcal{F}$. Assume that $E_{n+1}\subseteq E_n$, and $\bigcap_{n\in\mathbb{N}}E_n = \emptyset$. Thus $X_n\downarrow 0$ pointwise, and $\int_\Omega X_n\,d\mathbb{P} = \mathbb{P}(E_n)$. It is easy, however, to find sets $\{E_n\}$ satisfying the previous conditions, and yet $\mathbb{P}(E_n)=\infty$ for all n; take, for example, $E_n = (n,\infty)$ with $\mathbb{P}=\lambda\equiv$ LEBESGUE measure on the real line. In this case $X\equiv 0$, so the right-hand side of (4.1) is equal to zero; but the left-hand side is equal to infinity.

The difficulty is easily identified if we try to adapt the treatment of the special case in the proof of (4.1) to the Monotone Convergence Theorem 4.2. Indeed, for any decreasing sequence $\{E_n\}_{n\in\mathbb{N}}\subseteq\mathcal{F}$ we can write $E_1 = \bigcup_{k\in\mathbb{N}}F_k$ as the countable union of the disjoint subsets $F_k := E_k\setminus E_{k+1}$, thus $\mathbb{P}(F_k) = \mathbb{P}(E_k)-\mathbb{P}(E_{k+1})$ and by countable additivity

$$\mathbb{P}(E_1) = \sum_{k\in\mathbb{N}}\mathbb{P}(F_k) = \lim_{n\to\infty}\sum_{k=1}^{n-1}\mathbb{P}(F_k) = \lim_{n\to\infty}\Big(\mathbb{P}(E_1)-\mathbb{P}(E_n)\Big) = \mathbb{P}(E_1)-\lim_{n\to\infty}\mathbb{P}(E_n).$$

The additional hypothesis $\mathbb{P}(E_1)<\infty$ would allow us to subtract $\mathbb{P}(E_1)$ from both sides and conclude with

• the **Continuity from Above** property

$$\lim_{n\to\infty}\mathbb{P}(E_n) \;=\; 0\,,\quad \text{for any decreasing sequence } \left\{E_n\right\}_{n\in\mathbb{N}}\subset\mathcal{F} \text{ with } \bigcap_{n\in\mathbb{N}}E_n = \emptyset$$

under the hypothesis $\mathbb{P}(E_1)<\infty$ (see Exercise 4.15 in the same vein).

This is trivially satisfied, of course, if $\mathbb{P}(\Omega) < \infty$. In our special context, *this hypothesis is exactly the same as the requirement that all* $|X_n|$, $n \in \mathbb{N}$ *be dominated by an integrable function.*

Transported to the context of the proof of Proposition 2.1, the step of subtracting $\mathbb{P}(E_1)$ in the above simple case will require the *linearity property* (2.14) of the integral. This property is not evident from our definition of the integral in (2.8). (Nor, for that matter, is it completely evident that *linear combinations of measurable functions are measurable;* cf. Exercise 2.4.) However, assuming both these statements for the moment, we can prove the LEBESGUE Dominated Convergence Theorem 4.4 along the lines of the previous example.

Proof of Theorem 4.4: Since $|X_n| \leq Y$, we also have $|X| \leq Y$ for $X := \lim_{n\to\infty} X_n$. The sequence $\{2Y - |X - X_n|\}_{n\in\mathbb{N}} \subset \mathbb{L}^0_+$ converges to $2Y$, so by FATOU's Lemma and linearity of the integral:

$$\int_\Omega \big(2Y\big)\, \mathrm{d}\mathbb{P} \;\leq\; \liminf_{n\to\infty} \int_\Omega \big(2Y - |X - X_n|\big)\, \mathrm{d}\mathbb{P} = \int_\Omega (2Y)\, d\mathbb{P} \;-\; \limsup_{n\to\infty} \int_\Omega \big|X - X_n\big|\, \mathrm{d}\mathbb{P}\,.$$

It follows that $\limsup_{n\to\infty} \int_\Omega |X - X_n|\, \mathrm{d}\mathbb{P} \leq 0$, therefore $\lim_{n\to\infty} \int_\Omega |X - X_n|\, \mathrm{d}\mathbb{P} = 0$; so from (2.12) and linearity again, we get

$$\left| \int_\Omega X\, \mathrm{d}\mathbb{P} - \int_\Omega X_n\, \mathrm{d}\mathbb{P} \right| \;=\; \left| \int_\Omega \big(X - X_n\big)\, \mathrm{d}\mathbb{P} \right| \;\leq\; \int_\Omega \big|X - X_n\big|\, \mathrm{d}\mathbb{P} \;\longrightarrow\; 0\,, \quad \text{as } n \to \infty\,. \qquad \square$$

Proof of Proposition 2.1: We come now to the issue of the linearity property (2.14) for the integral. Let us begin by recalling that *the sum and the product of a finite number of measurable functions are measurable* (Exercise 2.4), and check first the linearity of the integral for *simple functions*.

Suppose that

$$Y = \sum_{j=1}^{M} y_j\, \mathbf{1}_{E_j}\,, \qquad Z = \sum_{k=1}^{N} z_k\, \mathbf{1}_{F_k}$$

are simple functions with values $\{y_1, \cdots, y_M\}$ and $\{z_1, \cdots, z_N\}$ respectively; here we have set $E_j := Y^{-1}(\{y_j\})$, $F_k := Z^{-1}(\{z_k\})$, and $E_i \cap E_j = \emptyset = F_k \cap F_\ell$ for $i \neq j$, $k \neq \ell$ without loss of generality. Then $\alpha Y + \beta Z$ is a simple function for any real constants α and β, with value $\alpha y_j + \beta z_k$ on the set $E_j \cap F_k$. Hence, by the finite additivity of the measure $\mathbb{P}$, we obtain the desired result

$$\int_\Omega (\alpha Y + \beta Z)\, \mathrm{d}\mathbb{P} \;=\; \sum_{j=1}^{M} \sum_{k=1}^{N} (\alpha y_j + \beta z_k) \mathbb{P}(E_j \cap F_k) \;=\; \alpha \sum_{j=1}^{M} y_j \mathbb{P}(E_j) + \beta \sum_{k=1}^{N} z_k \mathbb{P}(F_k)$$

$$= \;\alpha \int_\Omega Y\, \mathrm{d}\mathbb{P} \;+\; \beta \int_\Omega Z\, \mathrm{d}\mathbb{P}\,. \tag{4.8}$$

• The next step requires us to replace the supremum in the definition (2.8) by a more convenient limit. The key now is the approximation result of Proposition 4.1 below, according to which for any given nonnegative, measurable function $X \in \mathbb{L}_+$ there exists a monotone increasing sequence of nonnegative, *simple functions* $\{Y_n\}_{n\in\mathbb{N}}$ converging pointwise to X as in (4.1). In view of the Monotone Convergence Theorem, it follows then from (4.10) that

$$\mathbb{E}(X) \,\equiv\, \int_\Omega X \,\mathrm{d}\mathbb{P} \,=\, \lim_{n\to\infty} \uparrow \int_\Omega Y_n \,\mathrm{d}\mathbb{P} \,=\, \lim_{n\to\infty} \uparrow \mathbb{E}(Y_n)\,. \tag{4.9}$$

• Let now X and W be (*non-negative*, measurable) functions in $\mathbb{L}_+$, and α and β be positive constants. Let $\{Y_n\}_{n\in\mathbb{N}}$ and $\{Z_n\}_{n\in\mathbb{N}}$ be sequences of non-negative, simple functions, monotonically increasing to X and W, respectively.

Then $\{\alpha Y_n + \beta Z_n\}_{n\in\mathbb{N}}$ is a sequence of nonnegative, simple functions, monotonically increasing to the (nonnegative, measurable) function $\alpha X + \beta W \in \mathbb{L}_+$. The linearity property (2.14) follows now from (4.8) and (4.9):

$$\mathbb{E}\big(\alpha X + \beta W\big) \,=\, \lim_{n\to\infty} \uparrow \mathbb{E}\big(\alpha Y_n + \beta Z_n\big) \,=\, \lim_{n\to\infty} \uparrow \Big(\alpha\,\mathbb{E}\big(Y_n\big) + \beta\,\mathbb{E}\big(Z_n\big)\Big)$$

$$=\, \alpha \cdot \lim_{n\to\infty} \uparrow\, \mathbb{E}\big(Y_n\big) + \beta \cdot \lim_{n\to\infty} \uparrow \mathbb{E}\big(Z_n\big) \,=\, \alpha\,\mathbb{E}(X) + \beta\,\mathbb{E}(W)\,.$$

• Finally, we need to verify (2.14) for arbitrary integrable functions. Let us note that $\mathbb{L}^1$ is a real vector space: if X and W are *real-valued* integrable functions, then so is $H := \alpha X + \beta W$ for any real constants α and β. Recalling from (2.13) that $\mathbb{E}(cX) = c\,\mathbb{E}(X)$ is valid for every $c \in \mathbb{R}$, it becomes clear that we need verify (2.14) only for $\alpha = \beta = 1$.

We do this by separating X and W into their positive and negative parts $X^\pm$ and $W^\pm$. The details are straightforward; indeed, with $H = X + W$ we have

$$H^+ - H^- = X^+ - X^- + W^+ - W^-\,, \qquad \text{so that} \qquad H^+ + X^- + W^- = H^- + X^+ + W^+\,;$$

from what has already been shown, we obtain

$$\mathbb{E}(H^+) + \mathbb{E}(X^-) + \mathbb{E}(W^-) \,=\, \mathbb{E}(H^-) + \mathbb{E}(X^+) + \mathbb{E}(W^+)\,,$$

and this gives $\mathbb{E}(H) = \mathbb{E}(X) + \mathbb{E}(W)$. The proof of (2.14) is complete. □

Proposition 4.1. Approximation of measurable functions by simple functions: *For any given non-negative, measurable function $X \in \mathbb{L}_+$, there exists a monotone increasing sequence of non-negative, simple functions $\{Y_n\}_{n\in\mathbb{N}}$ converging pointwise to X, namely:*

$$0 \,\le\, Y_n(\omega) \,\le\, Y_{n+1}(\omega) \,\to\, X(\omega) \quad \textit{as } n \to \infty\,, \quad \textit{for every } \omega \in \Omega\,. \tag{4.10}$$

Proof: The simple functions $\{Y_n\}$ are easily constructed by partitioning the *range* of X (rather than its domain, as we are accustomed to from RIEMANN's theory [32] of integration...), say according to

[32] Pedestrian as this may appear today, it was quite a radical departure a hundred years ago: the RIEMANN theory of integration possesses no approximation result of such generality and power. In the definition of the RIEMANN integral,

dyadic rationals, as follows: For each $n \in \mathbb{N}$, we define $Y_n := n$ on the set $X^{-1}([n,\infty])$, and on the set $X^{-1}([0,n))$ we set

$$Y_n := k\,2^{-n} \quad \text{on} \quad X^{-1}\big(\big[\,k\,2^{-n}\,,\,(k+1)\,2^{-n}\big)\big)\,, \qquad \text{for } k = 0, \cdots, n2^n - 1\,. \tag{4.11}$$

We call this Y_n the "lower dyadic rational approximation of order n" for the measurable function X. The situation is messier to write down than to visualize, so please draw a picture.

Since

$$\big[k\,2^{-n}, (k+1)\,2^{-n}\big) = \bigcup_{\ell=2k}^{2k+1} \big[\,\ell\,2^{-(n+1)}, (\ell+1)\,2^{-(n+1)}\big)\,,$$

the only possible values of Y_{n+1} on this set are $2k\,2^{-(n+1)}$ and $(2k+1)\,2^{-(n+1)}$, both of which are at least as big as the value $k\,2^{-n}$ of Y_n on this set. Thus, we have $Y_{n+1} \geq Y_n$ on $X^{-1}([0,n))$.

Similarly, on $X^{-1}([n,\infty])$ we have $Y_n = n \leq Y_{n+1}$, so that $\{Y_n\}_{n\in\mathbb{N}}$ is a monotone increasing sequence of simple functions. Furthermore,

$$X - (1/2^n) \leq Y_n \leq X \qquad \text{holds on the set } \; X^{-1}([0,n))\,.$$

Any point $\omega \in \Omega$ with $X(\omega) < \infty$ belongs to some $X^{-1}([0,n))$ for n large enough, and thus satisfies $Y_n(\omega) \to f(\omega)$. But at points $\omega \in \Omega$ with $X(\omega) = \infty$ we have $Y_n(\omega) = n$ for all n, and thus again $Y_n(\omega) \to X(\omega)$, establishing (4.10). □

Sketch of Proof for Theorem 4.5: For the first claim, let us note that for every Borel set $B \in \mathcal{B}(\mathbb{R})$ we have $G := \mathfrak{h}^{-1}(B) \in \mathcal{G}$ (on the strength of the measurability of $\mathfrak{h}$), so

$$\big(\mathfrak{h}X\big)^{-1}(B) \,=\, \big\{\,\omega\in\Omega\,:\,\mathfrak{h}\big(X(\omega)\big)\in B\,\big\} \,=\, \big\{\,\omega\in\Omega\,:\,X(\omega)\in\mathfrak{h}^{-1}\big(B\big)\,\big\} \,=\, X^{-1}(G) \,\in\, \mathcal{F}\,,$$

this time on the strength of the measurability of X.

The identity

$$\mathbb{E}^{\mathbb{P}}(\mathfrak{h}\,X) = \int_{\mathfrak{S}} \mathfrak{h}\,\mathrm{d}\mu$$

is established easily for indicator functions $\mathfrak{h} = \mathbf{1}_G$ with $G \in \mathcal{G}$; for in this case it amounts to $\mathbb{P}(X^{-1}(G)) = \mu(G)$, which is just the definition of the induced measure μ. Next, this identity is established for simple functions, by the linearity of the integral; then for arbitrary nonnegative measurable $\mathfrak{h}$, by means of Proposition 4.1 and the monotone convergence theorem; and finally for $\mu-$integrable $\mathfrak{h}$, in the manner of the proof of Proposition 2.1.

We leave the details as a very important exercise. [33] □

---

one partitions the domain of the function, constructs rectangles to approximate its area, and then passes to the limit as the partitions become finer and finer. This presupposes regularity on the part of the function, essentially continuity, and a good topological structure on the domain. By contrast, the definition of the LEBESGUE integral partitions the *range* (image space) of the function, and places on the domain just a measure – whose definition is *devoid of any topological consideration*. No regularity is required on the part of the function – just measurability.

Twenty five years after its inception, the LEBESGUE approach turned out to be ideally suited for a mathematical foundation of Probability Theory, including the notion of expectation, which has to rely on as little topological structure as possible; as such it was exploited to the fullest by KOLMOGOROV (1933) in his foundational work.

[33] This kind of methodology will be invoked and used over and over again from now on; it should become second nature to us, to be able to argue along these lines.

4.4 Families of Sets

In the theories of measure and probability, we face often the following situation: We want to show that a certain property holds for all sets in a certain $\sigma-$algebra $\mathcal{G}$, but we can do this relatively easily only for sets in a subclass $\mathcal{D}$ of $\mathcal{G}$. *Under what conditions on $\mathcal{G}$ and $\mathcal{D}$ can we then ensure that the property holds on the larger class $\mathcal{G}$?*

The following two results introduce concepts and conditions that make such conclusions possible, in a fairly systematic way. They will be used rather extensively in what follows, so the reader will be well advised to think them through carefully – and forgive the strange terminology with which we have, rather unfortunately, been stuck.

Definition 4.1. Monotone Class: *A nonempty collection $\mathcal{M}$ of subsets of a non-empty space Ω is called a* Monotone Class, *if it is closed under countable increasing unions and under countable decreasing intersections.*

Definition 4.2. DYNKIN Systems: *A nonempty collection $\mathcal{D}$ of subsets of a non-empty space Ω is called a*

- $\pi-$system, *if it is closed under finite intersections; it is called a*
- $\lambda-$system, *if it contains Ω, if it is closed under countable increasing unions, and if we have $A \setminus B \in \mathcal{D}$ whenever $A \in \mathcal{D}$, $B \in \mathcal{D}$ and $B \subseteq A$.*

Clearly, every σ-algebra is a monotone class, as well as a $\pi-$system and a $\lambda-$system.

On the other hand, the intersection of an arbitrary collection of $\lambda-$systems is also a $\lambda-$system; so for any collection $\mathcal{A} \subseteq \mathcal{F}$ of subsets of Ω we can define $\lambda(\mathcal{A})$ as the intersection of all $\lambda-$systems that contain $\mathcal{A}$. This is the smallest $\lambda-$system that contains $\mathcal{A}$, and we have clearly $\mathcal{A} \subseteq \lambda(\mathcal{A}) \subseteq \sigma(\mathcal{A})$.

We have the following two basic results, Theorems 4.6 and 4.7 below, which are obvious analogues of each other.

Theorem 4.6. Monotone Class Theorem: *(i) The intersection of an arbitrary family of monotone classes is a monotone class; thus, for any family $\mathcal{E} \neq \emptyset$ of subsets of Ω, there is a smallest monotone class, denoted $m(\mathcal{E})$, which contains $\mathcal{E}$.*
(ii) If $\mathcal{E}$ is an algebra, then we have $\sigma(\mathcal{E}) = m(\mathcal{E})$.

Proof: (ii) Since every $\sigma-$algebra is a monotone class, it is clear that $\mathcal{E} \subseteq \mathcal{M} := m(\mathcal{E}) \subseteq \sigma(\mathcal{E})$ holds for *any* nonempty family $\mathcal{E}$ of subsets. We need to argue the reverse inclusion $\sigma(\mathcal{E}) \subseteq \mathcal{M}$ when $\mathcal{E}$ is an algebra. And for this, it is enough to show that $\mathcal{M}$ is a $\sigma-$algebra.

But actually, it is enough to argue that $\mathcal{M}$ *is an algebra.* For then, given any sequence $\{E_n\}_{n\in\mathbb{N}} \subseteq \mathcal{M}$, the sets $F_n := \bigcup_{j=1}^{n} E_j$, $n \in \mathbb{N}$ belong to $\mathcal{M}$; and since $\mathcal{M}$ is a monotone class, we deduce $\bigcup_{j\in\mathbb{N}} E_j = \bigcup_{n\in\mathbb{N}} F_n \in \mathcal{M}$, as this latter union is increasing.

To see that $\mathcal{M}$ is an algebra, let us introduce for any given set $G \in \mathcal{M}$ the collection of sets

$$\mathcal{C}(G) := \big\{F \in \mathcal{M} : F \setminus G,\ G \setminus F,\ F \cap G \quad \text{belong to } \mathcal{M}\big\}. \tag{4.12}$$

This collection contains the empty set, and the set G; it is also a monotone class. Furthermore,

$$F \in \mathcal{C}(G) \iff G \in \mathcal{C}(F) \tag{4.13}$$

holds for any $F \in \mathcal{M}$, $G \in \mathcal{M}$.

In particular, if $G \in \mathcal{E}$, we see that $F \in \mathcal{C}(G)$ holds for every $F \in \mathcal{E}$, because $\mathcal{E}$ is an algebra and $\mathcal{M} = m(\mathcal{E})$. In other words, we have then $\mathcal{E} \subseteq \mathcal{C}(G)$, thus also $\mathcal{M} \subseteq \mathcal{C}(G)$ because $\mathcal{C}(G)$ is a monotone class.

Therefore, for every $G \in \mathcal{E}$, $F \in \mathcal{M}$ we have $F \in \mathcal{C}(G)$, which amounts to $G \in \mathcal{C}(F)$ on account of (4.13). Consequently $\mathcal{E} \subseteq \mathcal{C}(F)$, thus also $\mathcal{M} \subseteq \mathcal{C}(F)$, hold for any $F \in \mathcal{M}$.

We conclude from this and (4.12), that for arbitrary $F \in \mathcal{M}$, $G \in \mathcal{M}$ the sets

$$F \setminus G\,, \quad G \setminus F\,, \quad F \cap G \qquad \text{belong to } \mathcal{M}\,;$$

and because $\Omega \in \mathcal{E} \subseteq \mathcal{M}$, we deduce that $\mathcal{M}$ is an algebra, as claimed. □

Theorem 4.7. DYNKIN System Theorem: *The following hold:*

(i) If $\mathcal{D}$ is both a $\pi-$system and a $\lambda-$system, then it is a σ-algebra (and vice-versa).

(ii) If $\mathcal{D}$ is a $\pi-$system, then we have $\sigma(\mathcal{D}) = \lambda(\mathcal{D})$.

In particular, any $\lambda-$system which contains a $\pi-$system, also contains the $\sigma-$algebra generated by that $\pi-$system.

Proof: *(i)* If $\mathcal{D}$ is both a $\pi-$system and a $\lambda-$system, then it is closed under complementation and finite unions. Indeed, for every sequence $\{E_n\}_{n\in\mathbb{N}} \subseteq \mathcal{D}$ we have $E_n^c = \Omega \setminus E_n \in \mathcal{D}$ and $E_1 \cup E_2 = \left(E_1^c \cap E_2^c\right)^c \in \mathcal{D}$.

To show that $\mathcal{D}$ is closed also under countable unions, just observe that $G_n := \bigcup_{j=1}^n E_j \in \mathcal{D}$ holds for every $n \in \mathbb{N}$, and that $G_n \uparrow \bigcup_{j\in\mathbb{N}} E_j =: G$; thus we have $G \in \mathcal{D}$ as well, since $\mathcal{D}$ is closed under countable increasing unions.

The reverse implication has already been noted, right after Definition 4.2.

(ii) Now let us show that we have $\lambda(\mathcal{D}) = \sigma(\mathcal{D})$ for any $\pi-$system $\mathcal{D}$. Because $\lambda(\mathcal{D}) \subseteq \sigma(\mathcal{D})$ holds trivially, as we have noted following Definition 4.2, it is enough to argue the reverse inclusion $\sigma(\mathcal{D}) \subseteq \lambda(\mathcal{D})$. But for this, and on account of part (i), we need only show that $\mathcal{A} := \lambda(\mathcal{D})$ is a $\pi-$system; that is, closed under pairwise intersections. We sketch this argument below, and refer to WILLIAMS (1991), page 194 for the details.

Consider first the class $\mathcal{A}_1 := \{\, A \in \mathcal{A} \,|\, A \cap B \in \mathcal{A}\,,\ \forall\, B \in \mathcal{D}\,\}$. Because $\mathcal{D}$ is a $\pi-$system we have $\mathcal{D} \subseteq \mathcal{A}_1$. We also can check that $\mathcal{A}_1$ is a $\lambda-$system, because so is $\mathcal{A}$. Since $\mathcal{A}$ is the smallest $\lambda-$system that contains $\mathcal{D}$, this shows that $\mathcal{A}_1 = \mathcal{A}$.

Next, let us look at the class $\mathcal{A}_2 := \{\, A \in \mathcal{A} \,|\, A \cap B \in \mathcal{A}\,,\ \forall\, B \in \mathcal{A}\,\}$ and deduce $\mathcal{D} \subseteq \mathcal{A}_2$ from $\mathcal{A}_1 = \mathcal{A}$. It is not hard to verify that $\mathcal{A}_2$ is a $\lambda-$system, therefore $\mathcal{A}_2 = \mathcal{A}$. We conclude that $\mathcal{A}$ is a $\pi-$system. □

Theorem 4.8. Determination of Measure: *Suppose $\mathcal{D}$ is a $\pi-$system, and two measures $\mathbb{P}$ and $\mathbb{Q}$ on $(\Omega, \sigma(\mathcal{D}))$ satisfy $\mathbb{P}(\Omega) = \mathbb{Q}(\Omega) < \infty$ and $\mathbb{P} \equiv \mathbb{Q}$ on $\mathcal{D}$.*

Then we have also $\mathbb{P} \equiv \mathbb{Q}$ on $\sigma(\mathcal{D})$.

Corollary 4.1. *If two probability measures agree on a collection of sets $\mathcal{D}$ which is closed under finite intersections, then they agree also on the σ-algebra $\sigma(\mathcal{D})$ generated by this collection.*

Proof of Theorem 4.8: The class $\mathcal{E} := \{ E \in \sigma(\mathcal{D}) \,|\, \mathbb{P}(E) = \mathbb{Q}(E) \}$ is a $\lambda-$system. Indeed, $\Omega \in \mathcal{E}$ by assumption; and if A, B with $B \subseteq A$ are in $\mathcal{E}$, we have

$$\mathbb{P}(A \setminus B) \,=\, \mathbb{P}(A) - \mathbb{P}(B) \,=\, \mathbb{Q}(A) - \mathbb{Q}(B) \,=\, \mathbb{Q}(A \setminus B)$$

because $\mathbb{P}$ and $\mathbb{Q}$ are finite measures (the finiteness assumption is crucial here), so $A \setminus B \in \mathcal{E}$; whereas for any increasing sequence $\{E_n\} \subseteq \mathcal{E}$ with $E := \bigcup_{n=1}^{\infty} E_n \in \mathcal{F}$ we have from (4.7)

$$\mathbb{P}(E) \,=\, \lim_n \uparrow \mathbb{P}(E_n) \,=\, \lim_n \uparrow \mathbb{Q}(E_n) \,=\, \mathbb{Q}(E)\,,$$

so $E \in \mathcal{E}$. By assumption $\mathcal{D} \subseteq \mathcal{E}$, and Theorem 4.7(ii) gives $\sigma(\mathcal{D}) = \lambda(\mathcal{D}) \subseteq \mathcal{E}$. □

Lemma 4.1. *On a given probability space $(\Omega, \mathcal{F}, \mathbb{P})$, suppose that $\mathcal{G}$ and $\mathcal{H}$ are sub-σ-algebras of $\mathcal{F}$ and that $\mathcal{I}$ and $\mathcal{J}$ are π-systems such that $\mathcal{G} = \sigma(\mathcal{I})$, $\mathcal{H} = \sigma(\mathcal{J})$.*

Then $\mathcal{G}$ and $\mathcal{H}$ are independent (Definition 2.5), if and only if $\mathcal{I}$ and $\mathcal{J}$ are independent.

Proof: Suppose $\mathcal{I}$ and $\mathcal{J}$ are independent; for any given $I \in \mathcal{I}$, the set functions

$$H \,\mapsto\, \mathbb{P}(I \cap H)\,, \qquad H \,\mapsto\, \mathbb{P}(I) \cdot \mathbb{P}(H)$$

agree on $\mathcal{J}$, are measures on $(\Omega, \mathcal{H})$, and have the same total mass $\mathbb{P}(I)$. By Theorem 4.7 they agree on $\mathcal{H} = \sigma(\mathcal{J})$, that is: $\mathbb{P}(I \cap H) = \mathbb{P}(I) \cdot \mathbb{P}(H)$ for every $H \in \mathcal{H}$.

Thus, for any given $H \in \mathcal{H}$, the set functions

$$G \,\mapsto\, \mathbb{P}(G \cap H)\,, \qquad G \,\mapsto\, \mathbb{P}(G) \cdot \mathbb{P}(H)$$

agree on $\mathcal{I}$; they are measures on $(\Omega, \mathcal{G})$ and have the same total mass $\mathbb{P}(H)$. By the same token as above, they they agree on $\mathcal{G} = \sigma(\mathcal{I})$; to wit, we have $\mathbb{P}(G \cap H) = \mathbb{P}(G) \cdot \mathbb{P}(H)$ for every $G \in \mathcal{G}$. □

Definition 4.3. Elementary Family: *A nonempty collection $\mathcal{G}$ of subsets of a non-empty space Ω is called* Elementary Family, *if it contains the empty set, is closed under finite intersections, and the complement of every $E \in \mathcal{G}$ can be written as a finite union $E^c = \bigcup_{j=1}^{n} F_j$ of pairwise disjoint sets $F_1, \cdots, F_n$ in $\mathcal{G}$.*

For instance, the collection of all half-open intervals $(a, b]$ with $-\infty \le a \le b \le \infty$ of the real line, is an elementary family. (When $b = \infty$, $a \in \mathbb{R}$, we interpret the interval as a half-line (a, ∞); when $a = b \in \mathbb{R}$, as the empty set.)

Exercise 4.13. Let $\mathcal{G}$ be an elementary family, and consider the collection $\mathcal{E}$ of finite disjoint unions of sets from $\mathcal{G}$. Show that $\mathcal{E}$ is an algebra.

Exercise 4.14. *The aim of this exercise is to show that* LEBESGUE *measure cannot be extended to all the subsets of the unit interval – that is, to the power set* $\mathcal{P}([0,1))$ *– without losing its characteristic property of shift-invariance.*

Let $\mathbb{P}$ denote LEBESGUE measure on $\big([0,1), \mathcal{B}([0,1))\big)$. For any given $a \in \mathbb{R}$, define the "shift mapping" $T_a(x) = x + a \mod 1$, for $x \in [0,1)$; for instance, $T_{0.9}(0.3) = 0.2$.

(i) Show that each T_a is measurable with respect to $\mathcal{B}([0,1))\big)$, and that LEBESGUE measure is invariant under T_a, that is, $\mathbb{P} \circ T_a^{-1} = \mathbb{P}$.

(ii) Now let $\mathcal{F}$ be *any* σ-algebra of subsets of [0,1) for which T_a is measurable, and assume $\mathbb{P}$ is defined and shift-invariant (as in (i) above) on all of $\mathcal{F}$. Show then, that any set $A \in \mathcal{F}$ with $\mathbb{P}(A) > 0$ contains points x, y such that $x \neq y$ and $x - y$ is irrational.
(*Hint:* Show that $\{T_q^{-1}(A),\, q \in \mathbf{Q} \cap [0,1)\}$ is a family of disjoint sets, and use part (i).)

(iii) Conclude that LEBESGUE measure cannot be extended to $\mathcal{P}([0,1))$, all the while retaining its characteristic property of shift-invariance.
(*Hint:* Consider the equivalence relation $x \sim y \iff x - y \in \mathbf{Q}$, and invoke the Axiom of Choice, to obtain a set $A \in \mathcal{P}([0,1))$ of representatives from each of the induced equivalence classes. This gives a decomposition $[0,1) = \bigcup_{r \in \mathbf{Q} \cap [0,1)} T_r^{-1}(A)$; what implications for $\mathbb{P}(A)$ does this decomposition have?)

4.5 Exercises

Exercise 4.15. Continuity of measure from above: Show that we have

$$\lim_{n\to\infty} \mathbb{P}(E_n) \;=\; \mathbb{P}\Big(\bigcap_{n\in\mathbb{N}} E_n\Big)\,, \tag{4.14}$$

whenever $\{E_n\}_{n\in\mathbb{N}} \subseteq \mathcal{F}$ is a monotone decreasing sequence of measurable sets: $E_{n+1} \subseteq E_n$ for every $n \in \mathbb{N}$, provided $\mathbb{P}(E_m) < \infty$ for some $m \in \mathbb{N}$.

In particular, this property holds for a probability measure.

Exercise 4.16. Approximating Measurable by Simple Functions: For every measurable function $X : \Omega \to \mathbb{R}$ there exists a sequence $\{Y_n\}_{n\in\mathbb{N}}$ of simple functions with $|Y_1| \leq |Y_2| \leq \ldots \leq |X|$; with $Y_n \longrightarrow X$ pointwise; and with $\sup_{\omega\in E} |Y_n(\omega) - X(\omega)| \longrightarrow 0$ as $n \to \infty$, for any set $E \in \mathcal{F}$ on which X is bounded.

Exercise 4.17. (i) Suppose μ is a measure on the BOREL subsets of the real line, finite on bounded intervals. We use it to define a function $F : \mathbb{R} \to \mathbb{R}$ via $F(0) := 0$ and

$$F(x) \;:=\; \mu\big((0,x]\big)\,,\;\; x > 0 \qquad \text{and} \qquad F(x) \;:=\; -\mu\big((x,0]\big)\,,\;\; x > 0\,.$$

This function is nondecreasing and right continuous, and for any real numbers $a < b$, we have

$$\mu((a,b]) = F(b) - F(a)\,,\quad \mu([a,b)) = F(b-) - F(a-)\,,\quad \mu(\{a\}) = F(a) - F(a-)\,, \tag{4.15}$$

$$\mu([a,b]) = F(b) - F(a-),\quad \mu((a,b)) = F(b-) - F(a)\,. \tag{4.16}$$

(ii) Given a random variable $X : \Omega \to \mathbb{R}$ on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, consider the induced measure $\mu_X = \mathbb{P} \cdot X^{-1}$. Show that the function

$$F_X(x) := \mu_X\big((-\infty, x]\big) = \mathbb{P}(X \leq x), \qquad x \in \mathbb{R}$$

is nondecreasing and right continuous, and satisfies $F_X(-\infty) = 0$, $F_X(\infty) = 1$ as well as the properties (4.15), (4.16) above.

Exercise 4.18. On a suitable probability space $(\Omega, \mathcal{F}, \mathbb{P})$ construct a sequence $\{Z_n\}$ of random variables with $\lim_{n\to\infty} Z_n(\omega) = -\infty$ for $\mathbb{P}-$a.e. $\omega \in \Omega$ and $\lim_{n\to\infty} \mathbb{E}(Z_n) = +\infty$.

Exercise 4.19. For any $X \in \mathbb{L}_+$ and $\{X_n\}_{n\in\mathbb{N}} \subset \mathbb{L}_+^*$ we have:

(i) $\mathbb{E}(X) = 0 \iff \mathbb{P}(X \neq 0) = 0$;

(ii) $\mathbb{E}\left(\sum_{n\in\mathbb{N}} X_n\right) = \sum_{n\in\mathbb{N}} \mathbb{E}(X_n)$;

(iii) $\mathbb{E}(X) = \sup_{\substack{Z\in\mathbb{L}_+^* \\ 0\leq Z\leq X}} \mathbb{E}(Z)$;

(iv) if $\mathbb{E}(X) < \infty$, then the set $A = \{\omega \in \Omega : X(\omega) = \infty\}$ is $\mathbb{P}-$null.

(v) if $\mathbb{P}$, $\mathbb{Q}$ are two measures with $\mathbb{P}(A) \leq \mathbb{Q}(A)$, $\forall\, A \in \mathcal{F}$, then

$$\int_\Omega X \, \mathrm{d}\mathbb{P} \leq \int_\Omega X \, \mathrm{d}\mathbb{Q}$$

holds for every $X \in \mathbb{L}_+$;

(vi) the mapping $\mathbb{Q} : \mathcal{F} \to [0, \infty]$ defined by $\mathbb{Q}(E) := \int_E X \, \mathrm{d}\mathbb{P}$ is a measure, and we have for every $Y \in \mathbb{L}_+$ the property

$$\int_\Omega Y \, \mathrm{d}\mathbb{Q} = \int_\Omega XY \, \mathrm{d}\mathbb{P}.$$

Exercise 4.20. For any integrable, real-valued functions X, Y and $\{X_n\}_{n\in\mathbb{N}}$ we have:

(i) $|\mathbb{E}(X)| \leq \mathbb{E}(|X|)$;

(ii) $\mathbb{E}(|X + Y|) \leq \mathbb{E}(|X|) + \mathbb{E}(|Y|)$;

(iii) $\int_A X \, \mathrm{d}\mathbb{P} = \int_A Y \, \mathrm{d}\mathbb{P}$, $\forall\, A \in \mathcal{F} \iff \mathbb{E}(|X - Y|) = 0 \iff \mathbb{P}(X \neq Y) = 0$;

(iv) if $\sum_{n\in\mathbb{N}} \mathbb{E}(|X_n|) < \infty$, then $\sum_{n\in\mathbb{N}} X_n$ converges $\mathbb{P}-$a.e. to an integrable function, with

$$\mathbb{E}\Big(\sum_{n\in\mathbb{N}} X_n\Big) = \sum_{n\in\mathbb{N}} \mathbb{E}(X_n).$$

Exercise 4.21. An Extended Monotone Convergence Theorem: On an arbitrary measure space $(\Omega, \mathcal{F}, \mathbb{P})$, suppose that the measurable functions $X_1 \leq \cdots \leq X_n \leq X_{n+1} \leq \cdots$ are integrable and monotonically increasing pointwise to $X = \lim_{n\to\infty} X_n$.

If $\sup_{n\in\mathbb{N}} \mathbb{E}(X_n) < \infty$, show that X is itself integrable, and that we have $\lim_{n\to\infty} \mathbb{E}(X_n) = \mathbb{E}(X)$.

Exercise 4.22. An Extended Dominated Convergence Theorem: Let $\{X_n\}$, $\{Y_n\}$ be functions in $\mathbb{L}^1$, such that

$$|X_n| \le Y_n \quad (\forall\, n \in \mathbb{N})\,, \qquad \lim_{n\to\infty} X_n \,=\, X\,, \qquad \lim_{n\to\infty} Y_n \,=\, Y$$

hold a.e. for some $X \in \mathbb{L}^0$ and $Y \in \mathbb{L}^1$. Assume also that $\lim_{n\to\infty} \mathbb{E}(Y_n) = \mathbb{E}(Y)$.

Show that we have then $X \in \mathbb{L}^1$, as well as

$$\lim_{n\to\infty} \mathbb{E}(X_n) \,=\, \mathbb{E}(X)\,, \qquad \lim_{n\to\infty} \mathbb{E}(|X_n - X|) \,=\, 0\,.$$

Exercise 4.23. Distribution-Determining Classes: On a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, show that the distribution of a random vector $X : \Omega \to \mathbb{R}^d$, namely the probability measure $\mu_X(\cdot) = \mathbb{P}(X \in \cdot\,)$, is determined uniquely by knowledge of the expectations

$$\mathbb{E}[\Psi(X)] = \int_\Omega \Psi(X(\omega))\, \mathrm{d}\mathbb{P}(\omega) = \int_{\mathbb{R}^d} \Psi\, \mathrm{d}\mu_X$$

for all bounded, continuous functions $\Psi : \mathbb{R}^d \to \mathbb{R}$.

Exercise 4.24. On a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, suppose the real-valued random variables X, Y are identically distributed and we have $X \ge Y$ a.e. Show then $X = Y$ a.e.

Exercise 4.25. Prove the Inclusion-Exclusion Formula of section 2.4.
(*Hint:* Proceed by induction. Alternatively, argue that

$$\mathbf{1}_{\bigcup_{i=1}^n E_i} \,=\, 1 - \prod_{i=1}^n (1 - \mathbf{1}_{E_i})\,;$$

then expand the right-hand side and take expectations.)

Exercise 4.26. Regular Measure: Suppose that Ω is a metric space, and $\mathcal{F}$ a σ-algebra of its subsets that contains the Borel sets: $\mathcal{B}(\Omega) \subseteq \mathcal{F}$.

A measure $\mu : \mathcal{F} \to [0, \infty]$ is called *regular*, if for every $E \in \mathcal{F}$ and $\varepsilon > 0$ there exist an open set G and a closed set F such that $F \subseteq E \subseteq G$ and $\mu(G \setminus F) < \varepsilon$.

(i) If μ as above is regular, then an arbitrary set $D \in \mathcal{F}$ is "very near a Borel set", in the sense that we have

$$B_1 \subseteq D \subseteq B_2, \quad \text{for some } B_1,\, B_2 \text{ in } \mathcal{B}(\Omega) \text{ with } \mu(B_2 \setminus B_1) = 0$$

as well as

$$D = E \cup F, \quad \text{for some } E \in \mathcal{B}(\Omega),\, F \subseteq B \in \mathcal{B}(\Omega) \text{ with } \mu(B) = 0.$$

(i) *If a measure μ on $\mathcal{B}(\Omega)$ is finite on bounded* BOREL *sets, then it is regular.*

(*Hint:* Start with the finite case $\mu(\Omega) < \infty$. Consider the collection $\mathcal{E}$ of subsets E of Ω, such that for each ε there exist a closed set F_ε and an open set G_ε with $F_\varepsilon \subseteq E \subseteq G_\varepsilon$, $\mu\,(G_\varepsilon \setminus F_\varepsilon) < \varepsilon$. Argue that $\mathcal{E}$ is a σ-algebra that contains the closed sets, hence $\mathcal{B}(\Omega) \subseteq \mathcal{E}$. Then try to remove the assumption $\mu(\Omega) < \infty$.)

5 Essentials

We present in this chapter the basic inequalities of the subject; an introduction to the theory of LEBESGUE ($\mathbb{L}^p$) spaces of integrable functions; the theory and properties of product measure; as well as the LEBESGUE Decomposition and RADON-NIKODÝM theorems.

5.1 The ČEBYŠEV Inequality

Given a measure space $(\Omega, \mathcal{F}, \mathbb{P})$, consider measurable functions $X : \Omega \to \mathbb{R}$ and $h : \mathbb{R} \to [0, \infty)$ as well as a set $A \in \mathcal{B}(\mathbb{R})$. We have

$$\big(\inf_{x \in A} h(x)\big) \cdot \mathbb{P}\big(X \in A\big) \;\leq\; \mathbb{E}\big[\, h(X)\, \mathbf{1}_A(X)\,\big] \;\leq\; \mathbb{E}\big[\, h(X)\,\big]\,. \tag{5.1}$$

In particular, if $A = \{x \in \mathbb{R} \mid |x| \geq c\}$ for some $c > 0$ with $h(c) > 0$, and h is evenly symmetric (i.e., $h(-x) = h(x)$ for all $x \in \mathbb{R}$), and nondecreasing on $[0, \infty)$, we have the ČEBYŠEV *Inequality*

$$\boxed{\mathbb{P}(|X| \geq c) \;\leq\; \mathbb{E}\big[\, h(X)\,\big] \,/\, h(c)\,.} \tag{5.2}$$

For instance, if X is an integrable random variable, then from (5.2) with $h(c) = |c|$ we get

$$\mathbb{P}\big(|X| \geq c\big) \;\leq\; \frac{\mathbb{E}(|X|)}{c}\,, \qquad \forall\; c > 0\,. \tag{5.3}$$

If on the other hand Y is a square-integrable random variable, then from (5.2) with $X = Y - \mathbb{E}(Y)$ and $h(c) = c^2$ we get

$$\mathbb{P}\big(|Y - \mathbb{E}(Y)| \geq c\big) \;\leq\; \frac{\mathrm{Var}(Y)}{c^2}\,, \qquad \forall\; c > 0\,. \tag{5.4}$$

Exercise 5.1. S. BERNŠTEIN's proof of the WEIERSTRASS Approximation Theorem: For any continuous function $f : [0, 1] \to \mathbb{R}$, there exists a sequence of *polynomials* $\{B_n(\cdot)\}_{n \in \mathbb{N}}$ that converge to f uniformly over $[0, 1]$:

$$\sup_{0 \leq x \leq 1} \big|B_n(x) - f(x)\big| \;\longrightarrow\; 0\,, \qquad \text{as } n \to \infty\,.$$

In fact, one can take as such the so-called "BERNŠTEIN *polynomials*"

$$B_n(x) \;=\; \sum_{k=0}^{n} f(k/n) \cdot \binom{n}{k}\; x^k (1-x)^{n-k}\,, \qquad 0 \leq x \leq 1\,,\;\; n \in \mathbb{N}\,. \tag{5.5}$$

5.2 The HÖLDER and MINKOWSKI Inequalities

For any given real number $p > 0$, let us denote by $\mathbb{L}^p$ the set of measurable, real-valued functions X on the measure space $(\Omega, \mathcal{F}, \mathbb{P})$ with $|X|^p \in \mathbb{L}^1$, or equivalently

$$||X||_p \;:=\; \left(\int_\Omega |X|^p \, \mathrm{d}\mathbb{P} \right)^{1/p} \;< \infty\,. \tag{5.6}$$

We shall say that a sequence $\{X_n\}_{n\in\mathbb{N}}$ of functions in $\mathbb{L}^p$ *converges to some function* X *in* $\mathbb{L}^p$, if $\lim_{n\to\infty} ||X_n - X||_p = 0$.

The elementary properties

$$|X+Y|^p \le (2\cdot\max(|X|,|Y|))^p \le 2^p\cdot(|X|^p+|Y|^p)\ , \qquad ||\alpha X||_p = |\alpha|\,||X||_p$$

for real numbers α and $p>0$, show that each $\mathbb{L}^p$ is a real vector space. In fact, let us also note the inequality

$$|X+Y|^p \ \le\ |X|^p + |Y|^p\,, \qquad \text{for } 0<p<1\,.$$

On the other hand, the "triangle inequality" (5.8) below shows that, for $1\le p<\infty$, the quantity $||\cdot||_p$ of (5.6) is a norm on $\mathbb{L}^p$.

Important Remark: We continue to employ here and in the sequel the "usual convention" of identifying functions that are equal $\mathbb{P}-$a.e. on Ω; for instance, we identify $X=\mathbf{1_Q}$ with $Y=0$ on the real line with LEBESGUE measure. Thus, we are (tacitly) treating $\mathbb{L}^p$ as a space of *equivalence classes* of functions, rather than as a space of functions.

HÖLDER INEQUALITY: *For any* $p\in(1,\infty)$, *define* q *by* $(1/p)+(1/q)=1$; *then for any measurable, real-valued functions* $X,\,Y$ *we have*

$$\boxed{\ ||XY||_1 \le ||X||_p\cdot||Y||_q\ .\ } \tag{5.7}$$

If $X\in\mathbb{L}^p$ *and* $Y\in\mathbb{L}^q$, *this shows* $XY\in\mathbb{L}^1$, *and in this case (5.7) holds as equality* if and only if *there exist real constants* $\alpha,\,\beta$ *such that* $\alpha\,\beta\neq 0$ *and* $\alpha|X|^p=\beta|Y|^q$, $\mathbb{P}-$a.e.

For $p=2$, the inequality (5.7) is known as the **CAUCHY-SCHWARZ inequality.**

MINKOWSKI INEQUALITY: *For any* $p\in[1,\infty)$, *we have for every* $X,\,Y\,\in\mathbb{L}^p$ *the* triangle inequality

$$\boxed{\ ||X+Y||_p\ \le\ ||X||_p+||Y||_p\ .\ } \tag{5.8}$$

Exercise 5.2. *The triangle inequality (5.8) fails for* $p\in(0,1)$.
(*Hint:* Justify the inequality

$$(a+b)^p < a^p+b^p$$

for $a>0$, $b>0$, $0<p<1$, and write it with $a=(\mathbb{P}(E))^{1/p}$, $b=(\mathbb{P}(F))^{1/p}$ for any two disjoint measurable sets $E,\,F$ of positive measure.)

Exercise 5.3. An elementary YOUNG inequality: For $a\ge 0$, $b\ge 0$, $0<\lambda<1$ we have

$$a^\lambda\cdot b^{1-\lambda}\ \le\ \lambda a+(1-\lambda)b\,, \tag{5.9}$$

with equality iff $a=b$. (*Hint:* The function $\xi(u)=u^\lambda-u\lambda$ attains its maximum, namely $1-\lambda$, over the half-line $[0,\infty)$, at $u=1$.)

We can express this, equivalently, as the YOUNG inequality

$$xy\le\frac{x^p}{p}+\frac{y^q}{q}\,, \qquad \frac{1}{p}+\frac{1}{q}=1 \tag{5.10}$$

valid for all $x\ge 0$, $y\ge 0$ and with equality iff $x^p=y^q$; just take $\lambda=1/p$, $a=x^p$, $b=y^q$.

PROOF OF (5.7) : The inequality is obvious when its right-hand side $||X||_p \cdot ||Y||_q$ vanishes, so let us assume $||X||_p > 0,\ ||Y||_q > 0$. Then we can read the inequality (5.10) of Exercise 5.3 with $x = |X(\omega)|/||X||_p\,,\ y = |Y(\omega)|/||Y||_q\,$, to wit

$$\frac{|X(\omega)\,Y(\omega)|}{||X||_p\,||Y||_q} \le \frac{|X(\omega)|^p}{p\,\mathbb{E}(|X|^p)} + \frac{|Y(\omega)|^q}{q\,\mathbb{E}(|Y|^q)}, \quad \text{for } \omega \in \Omega$$

(with equality, iff $\mathbb{E}(|Y|^q) \cdot |X(\omega)|^p = \mathbb{E}(|X|^p) \cdot |Y(\omega)|^q$ holds for $\mathbb{P}-$a.e. $\omega \in \Omega$) . Integrating over Ω with respect to $\mathbb{P}$, we obtain $(||XY||_1) \,/\, (||X||_p\,||Y||_q\,) \le (1/p) + (1/q) = 1$.

PROOF OF (5.8) : The inequality is quite clear for $p = 1$, as well as when $X + Y = 0,\ \mathbb{P}-$a.e. Now for $p > 1$ and $\mathbb{P}(X + Y \neq 0) > 0$, we start by writing

$$|X + Y|^p \ \le\ |X| \cdot |X + Y|^{p-1} + |Y| \cdot |X + Y|^{p-1}\,;$$

integrating with respect to $\mathbb{P}$ and then applying HÖLDER's inequality to the right-hand side, we obtain

$$\begin{aligned}\mathbb{E}\,(|X + Y|^p) \ &\le\ ||X||_p \cdot ||(|X + Y|)^{p-1}||_q + ||Y||_p \cdot ||(|X + Y|)^{p-1}||_q \\ &\le\ \big(||X||_p + ||Y||_p\big) \cdot \Big(\mathbb{E}\,\big(|X + Y|^{q(p-1)}\big)\Big)^{1/q} = \big(||X||_p + ||Y||_p\big) \cdot \Big(\mathbb{E}\,(|X + Y|^p)\Big)^{1/q}\end{aligned}$$

whence $\big(\mathbb{E}\,(|X + Y|^p)\big)^{1-(1/q)} = ||X + Y||_p \le ||X||_p + ||Y||_p\,.$ □

• **The Space of Essentially Bounded Functions:** We can also define a space $\mathbb{L}^\infty\,,$ as the set of all (equivalence classes of) measurable functions $X : \Omega \to \mathbb{R}$ which are essentially bounded, in the sense that the *essential least-upper-bound*

$$\begin{aligned}||X||_\infty \ &:=\ \inf\{\,a \ge 0\,:\,|X(\omega)| \le a \ \text{ for } \mathbb{P} - \text{a.e. } \omega \in \Omega\,\} \\ &=\ \sup\{a \ge 0\,:\,\mathbb{P}\,(\{\omega \in \Omega\,:\,|X(\omega)| > a\}) > 0\} \qquad (5.11)\end{aligned}$$

is finite: $||X||_\infty < \infty\,.$ Under the usual convention, it is straightforward to check that $\mathbb{L}^\infty$ is a real vector space with $||\cdot||_\infty$ as its norm.

The essential least-upper-bound ignores sets of measure zero; for instance, if $X(\omega) = 1$ for rational $\omega \in \mathbb{R}$ and $X(\omega) = 0$ otherwise, then $||X||_\infty = 0$ but $\sup_{\omega\in\Omega}|X(\omega)| = 1\,.$

Exercise 5.4. (i) Note that the infimum in (5.11) is actually attained.
(ii) Suppose $X \in \mathbb{L}^\infty\,.$ Then $|X(\omega)| \le ||X||_\infty$ for $\mathbb{P}-$a.e. $\omega \in \Omega\,;$ and for every $0 < a < ||X||_\infty$ there exists a set $E \in \mathcal{F}$ with $\mathbb{P}(E) > 0$ such that $|X(\omega)| > a\,,\ \forall\,\omega \in E\,.$
(iii) For any $X \in \mathbb{L}^1,\ Y \in \mathbb{L}^\infty,\ W \in \mathbb{L}^\infty$ and $\{Y_n\}_{n\in\mathbb{N}} \subseteq \mathbb{L}^\infty$, we have the analogues

$$||XY||_1 \ \le\ ||X||_1 \cdot ||Y||_\infty\,, \qquad ||Y + W||_\infty \ \le\ ||Y||_\infty + ||W||_\infty$$

of the Hölder inequality and of the MINKOWSKI inequality, respectively.
(iv) For some set $E \in \mathcal{F}$ with $\mathbb{P}(E^c) = 0$, we also have the equivalence

$$\lim_{n\to\infty} ||Y_n - Y||_\infty = 0 \quad \Longleftrightarrow \quad \lim_{n\to\infty} \left(\sup_{\omega\in E} |Y_n(\omega) - Y(\omega)|\right) \ =\ 0\,.$$

In other words, convergence in $\mathbb{L}^\infty$ is uniform convergence outside a set of measure zero.

5.3 The JENSEN Inequality

A function $F : (a, b) \to \mathbb{R}$ is called *convex*, if for any $y_1, \ldots y_K$ in (a, b) and any $\lambda_1, \ldots, \lambda_K$ in $[0, 1]$ with $\lambda_1 + \cdots + \lambda_K = 1$ and any $K \in \mathbb{N}$, we have

$$F\Big(\sum_{k=1}^{K} \lambda_k y_k\Big) \le \sum_{k=1}^{K} \lambda_k F(y_k) .$$

In particular, if we interpret $\{y_1, \ldots y_K\}$ as the range of a simple function X on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and λ_k as $\mathbb{P}\big(X^{-1}(\{y_k\})\big)$, then the inequality reads

$$F\big(\mathbb{E}(X)\big) \le \mathbb{E}\big(F(X)\big) .$$

This is actually valid for any integrable function X, as the following result demonstrates.

JENSEN INEQUALITY: *Suppose that $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space, that $X : \Omega \to (a, b)$ is in $\mathbb{L}^1$, and that $F : (a, b) \to \mathbb{R}$ is a convex function, for some $-\infty \le a < b \le \infty$. We have then*

$$\boxed{F\big(\mathbb{E}(X)\big) \le \mathbb{E}\big(F(X)\big) .} \tag{5.12}$$

Proof: From the classical theory of convex functions of a real variable (e.g., ROBERTS & VARBERG (1973)) we know that, for each $\xi \in (a, b)$, there exists an affine function $L(x) = \alpha\, x + \beta$, $x \in (a, b)$ with $L(\cdot) \le F(\cdot)$ on (a, b) and $L(\xi) = F(\xi)$. We shall take $\xi = \mathbb{E}(X)$.

Let us notice that

$$\mathbb{E}\big(F^-(X)\big) \le \mathbb{E}\big(L^-(X)\big) \le |\alpha|\, \mathbb{E}\big(|X|\big) + |\beta| < \infty ,$$

so $\mathbb{E}\big(F(X)\big)$ is well defined. Moreover, the monotonicity of the expectation and the affinity of $L(\cdot)$ give the comparisons $\mathbb{E}\big(F(X)\big) \ge \mathbb{E}\big(L(X)\big) = L\big(\mathbb{E}(X)\big) = F\big(\mathbb{E}(X)\big)$. □

Exercise 5.5. LYAPUNOV Inequality: If X is a random variable on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and $0 < p < q \le \infty$, then $||X||_p \le ||X||_q$. In particular, $\mathbb{L}^q \subseteq \mathbb{L}^p$.

Exercise 5.6. LEGENDRE Transform Representation: Fix an arbitrary $p \in (1, \infty)$ and its "dual" q via $(1/p) + (1/q) = 1$. Show that the norm of every function $X \in \mathbb{L}^p$ admits the LEGENDRE Transform Representation

$$\big\|X\big\|_p^2 = \sup_{Y \in \mathbb{L}^q} \Big(2\,\big\|X\,Y\big\|_1 - \big\|Y\big\|_q^2\Big) .$$

Exercise 5.7. BERNŠTEIN polynomials (cont'd): In the context of Exercise 5.1 assume now, in addition to the assumptions made there, that the function $f(\cdot)$ satisfies the LIPSCHITZ condition $|f(x) - f(y)| \le K\,|x - y|$, $\forall\, x, y \in [0, 1]$ for some $K \in (0, \infty)$.

Show then that the BERNŠTEIN polynomials of (5.13) satisfy

$$\sup_{0 \le x \le 1} |B_n(x) - f(x)| \le \frac{K}{2\sqrt{n}} , \quad \text{for all } n \in \mathbb{N}.$$

5.4 Product Measure, TONELLI and FUBINI

Consider now *two* measurable spaces $(\Omega_1, \mathcal{F}_1)$, $(\Omega_2, \mathcal{F}_2)$ and introduce the Cartesian product

$$\Omega := \Omega_1 \times \Omega_2 := \big\{(\omega_1, \omega_2) : \omega_1 \in \Omega_1 \ \omega_2 \in \Omega_2\big\}.$$

The collection of "measurable rectangles"

$$\mathcal{R} := \big\{E_1 \times E_2 : E_1 \in \mathcal{F}_1,\ E_2 \in \mathcal{F}_2\big\}$$

is then both an elementary family and a π-system (Definitions 4.3 and 4.2, respectively): contains the empty set, is closed under finite intersections, and the complement of every rectangle in $\mathcal{R}$ is a finite union of pairwise-disjoint rectangles in $\mathcal{R}$. We define the *product* $\sigma-$*algebra* $\mathcal{F}_1 \otimes \mathcal{F}_2$ as the $\sigma-$algebra generated by $\mathcal{R}$, and call the resulting measurable space $(\Omega, \mathcal{F}) \equiv (\Omega_1 \times \Omega_2, \mathcal{F}_1 \otimes \mathcal{F}_2)$ the *product* of the given two measurable spaces. [34]

On this product measurable space $(\Omega, \mathcal{F})$, we would like to define a so-called *product measure* $\mathbb{P} \equiv \mathbb{P}_1 \otimes \mathbb{P}_2$, with the property

$$\mathbb{P}(E_1 \times E_2) = \mathbb{P}_1(E_1) \cdot \mathbb{P}_2(E_2), \quad \forall\ E_1 \in \mathcal{F}_1,\ E_2 \in \mathcal{F}_2\,. \tag{5.13}$$

We do this by a method, which uses integration theory to obtain an *integral representation* of this product measure (Theorem 5.1 below); then exploits this representation to study the properties of integration with respect to this measure (the TONELLI-FUBINI Theorems 6.2, 6.3).

In order to describe these results, let us introduce the "section notation"

$$E_{\omega_i} := \{\omega_j \in \Omega_j : (\omega_1, \omega_2) \in E\} \quad \text{and} \quad X_{\omega_i}(\omega_j) := X(\omega_1, \omega_2)\,,\ \ \omega_j \in \Omega_j \quad (j \neq i) \tag{5.14}$$

for subsets E of the product space Ω and for functions $X : \Omega \to \mathbb{R}$, with $\omega_i \in \Omega_i$ fixed. Clearly, for a rectangle $E = E_1 \times E_2$ with $E_i \in \mathcal{F}_i$ $(i = 1, 2)$ we have: $E_{\omega_1} = E_2$ (resp., $\emptyset$) for $\omega_1 \in E_1$ (resp., for $\omega_1 \notin E_1$).

The following simple fact can be checked rather easily: *For any given set* $E \in \mathcal{F}_1 \otimes \mathcal{F}_2$, *the section* E_{ω_i} *is in* $\mathcal{F}_j$, *for* $j \neq i$; *and if the mapping* X *is* $(\mathcal{F}_1 \otimes \mathcal{F}_2)-$*measurable, then the function* $\omega_j \mapsto X_{\omega_i}(\omega_j)$ *is* $\mathcal{F}_j-$*measurable, for* $j \neq i$.

Indeed, the collection $\mathcal{G}$ of subsets E of $\Omega_1 \times \Omega_2$ with

$$E_{\omega_1} \in \mathcal{F}_2 \ \text{ for all } \omega_1 \in \Omega_1\,, \qquad \text{and} \qquad E_{\omega_2} \in \mathcal{F}_1 \ \text{ for all } \omega_2 \in \Omega_2$$

contains all measurable rectangles $E = E_1 \times E_2$ $(E_1 \in \mathcal{F}_1\,,\ E_2 \in \mathcal{F}_2\,)$, and is a σ-algebra:

$$\left(\bigcup_{n\in\mathbb{N}} E^{(n)}\right)_{\omega_i} = \bigcup_{n\in\mathbb{N}} \left(E^{(n)}\right)_{\omega_i}, \qquad (E^c)_{\omega_i} = (E_{\omega_i})^c$$

[34] The product $\sigma-$algebra $\mathcal{B}(\mathbb{R}) \otimes \mathcal{B}(\mathbb{R})$ coincides with the $\sigma-$algebra $\mathcal{B}(\mathbb{R}^2)$. There are sets in this $\sigma-$algebra, whose projections on the axes are not BOREL subsets of the real line. The discovery of such sets by M. SUSLIN, who disproved along the way an erroneous statement to the contrary by none other than H. LEBESGUE himself, led to the development of Descriptive Set Theory. Take a look at the fascinating book by GRAHAM & KANTOR (2009) for a glimpse at the history of this subject and the flow of ideas between the French and Russian schools in the early 20th century.

hold for $i = 1, 2$; consequently, $\mathcal{G} \supseteq \mathcal{F} \otimes \mathcal{F}_2$. The second claim then follows, since we have $(X_{\omega_i})^{-1}(B) = (X^{-1}(B))_{\omega_i}$ for $i = 1, 2$. □

Furthermore, the mapping

$$\omega_i \longmapsto g_i(\omega_i) \equiv \mathbb{P}_j(E_{\omega_i}) \text{ is } \mathcal{F}_i - \text{measurable, for } j \neq i\,, \tag{5.15}$$

at least when both spaces are σ-finite, as we shall see below (proof of Theorem 5.1).

To illustrate the integral representation methods for constructing the product measure, let us recall from elementary calculus the computation of the area of the unit disc by means of a *single* integral.

Exercise 5.8. Let $\Omega_i = \mathbb{R}$, $\mathcal{F}_i = \mathcal{B}(\mathbb{R})$ with $\mathbb{P}_i = \lambda$ LEBESGUE measure ($i = 1, 2$) and consider the unit circle $E = \{(\omega_1, \omega_2) \,|\, \omega_1^2 + \omega_2^2 \leq 1\} \in \mathcal{B}(\Omega)$ in the product-space $\Omega = \mathbb{R}^2$. Then in the notation of (5.14) and (5.15) we have:

$$E_{\omega_1} = \left\{\omega_2 \in \Omega_2 \,:\, |\omega_2| \leq \sqrt{1 - \omega_1^2}\,\right\}, \qquad g_1(\omega_1) = \mathbb{P}_2(E_{\omega_1}) = 2\sqrt{1 - \omega_1^2} \qquad \text{for } |\omega_1| \leq 1\,;$$

and $E_{\omega_1} = \emptyset$, $g_1(\omega_1) = \mathbb{P}_2(E_{\omega_1}) = 0$ for $|\omega_1| > 1$. The resulting function $g_1(\cdot)$ is clearly measurable, and its integral is the area of the unit disc

$$\begin{aligned}\int_{\Omega_1} g_1 \,\mathrm{d}\mathbb{P}_1 \;&=\; \int_{-1}^{1} 2\sqrt{1 - \omega_1^2}\,\mathrm{d}\omega_1 \;=\; \int_0^1 4\sqrt{1 - t^2}\,\mathrm{d}t \\ &= \int_0^{\pi/2} 4\sin^2(\theta)\,\mathrm{d}\theta \;=\; \int_0^{\pi/2} 2\big[1 - \cos(\theta)\big]\,\mathrm{d}\theta = \pi\,.\end{aligned}$$

The calculation of this example can be generalized vastly, and in a way that leads directly to a measure on the product σ-algebra with the desired property (5.13).

Theorem 5.1. PRODUCT MEASURE: *If the component measure spaces are σ-finite, then the claim of (5.15) holds, and the set-function*

$$\mathbb{P}(E) := \int_{\Omega_1} g_1 \,\mathrm{d}\mathbb{P}_1 = \int_{\Omega_1} \mathbb{P}_2(E_{\omega_1})\,\mathrm{d}\mathbb{P}_1(\omega_1)\,, \quad E \in \mathcal{F} \tag{5.16}$$

is a σ-finite measure on the product σ-algebra $\mathcal{F} = \mathcal{F}_1 \otimes \mathcal{F}_2$. This measure satisfies

$$\mathbb{P}(E) = \int_{\Omega_2} g_2 \,\mathrm{d}\mathbb{P}_2 = \int_{\Omega_2} \mathbb{P}_1(E_{\omega_2})\,\mathrm{d}\mathbb{P}_2(\omega_2)\,, \qquad \forall\; E \in \mathcal{F}\,, \tag{5.17}$$

thus also the property (5.13); and is the unique *measure on $\mathcal{F}$ with the property (5.13).*

The measure of Theorem 5.1 is denoted $\mathbb{P} = \mathbb{P}_1 \otimes \mathbb{P}_2$ and is called **product measure** of $\mathbb{P}_1$, $\mathbb{P}_2$ on $\mathcal{F}_1 \otimes \mathcal{F}_2$. Clearly, $\mathbb{P}(E) = 0$, if and only if: $\mathbb{P}_j(E_{\omega_i}) = 0$ for $\mathbb{P}_i-$a.e. $\omega_i \in \Omega_i$ $(i \neq j)$. And $\mathbb{P}$ is a probability measure, if both $\mathbb{P}_1$, $\mathbb{P}_2$ are probability measures.

The following two fundamental results describe the properties of integration with respect to this product measure; first for positive, and then for general, real-valued functions on Ω.

Theorem 5.2. **TONELLI:** *In the context of Theorem 5.1, let* $X : \Omega \to [0,\infty)$ *be* $\mathcal{F}-$*measurable. Then the functions*

$$\omega_i \;\mapsto\; h_i(\omega_i) \;:=\; \int_{\Omega_j} X_{\omega_i}\,\mathrm{d}\mathbb{P}_j \;=\; \int_{\Omega_j} X_{\omega_i}(\omega_j)\,\mathrm{d}\mathbb{P}_j(\omega_j) \quad are\ \mathcal{F}_i-\text{measurable}, \tag{5.18}$$

for $1 \le i \ne j \le 2$*, and we have*

$$\int_\Omega X\,\mathrm{d}\mathbb{P} \;=\; \int_{\Omega_1} h_1\,\mathrm{d}\,\mathbb{P}_1 \;=\; \int_{\Omega_2} h_2\,\mathrm{d}\mathbb{P}_2\,; \tag{5.19}$$

or, a bit more suggestively,

$$\int\int_{\Omega_1\times\Omega_2} X(\omega_1,\omega_2)\;\mathrm{d}(\mathbb{P}_1\otimes\mathbb{P}_2)(\omega_1,\omega_2) \;=\; \int_{\Omega_1}\left(\int_{\Omega_2} X_{\omega_1}(\omega_2)\,\mathrm{d}\mathbb{P}_2(\omega_2)\right)\mathrm{d}\mathbb{P}_1(\omega_1)$$

$$=\; \int_{\Omega_2}\left(\int_{\Omega_1} X_{\omega_2}(\omega_1)\,\mathrm{d}\mathbb{P}_1(\omega_1)\right)\mathrm{d}\mathbb{P}_2(\omega_2)\,.$$

Theorem 5.3. **FUBINI:** *In the context of Theorem 5.1, consider a function* $X \in \mathbb{L}^1(\Omega,\mathcal{F},\mathbb{P})$*, that is, integrable on the product-space. Then the section* $X_{\omega_i}(\cdot)$ *belongs to* $\mathbb{L}^1(\Omega_j,\mathcal{F}_j,\mathbb{P}_j)$ *for* $\mathbb{P}_i-$*a.e.* $\omega_i \in \Omega_i$ $(1 \le i \ne j \le 2)$*, and*

the function h_i *of (5.18) is integrable, that is, belongs to* $\mathbb{L}^1(\Omega_i,\mathcal{F}_i,\mathbb{P}_i)$*,*

for $i = 1, 2$*. Furthermore, the identities of (5.19) hold.*

Proof of Theorem 5.1 : Consider first the finite case $\mathbb{P}_1(\Omega_1) + \mathbb{P}_2(\Omega_2) < \infty$. We need to verify that the family

$$\mathcal{M} \;:=\; \big\{E \subseteq \Omega_1\times\Omega_2 \;:\; \omega_1 \mapsto \mathbb{P}_2(E_{\omega_1}) = g_1(\omega_1) \;\text{ is }\; \mathcal{F}_1-\text{measurable}\big\}$$

contains all product-measurable sets: $\mathcal{M} \supseteq \mathcal{F}_1\otimes\mathcal{F}_2$.

Indeed, $\mathcal{M}$ contains the elementary family $\mathcal{R}$ of measurable rectangles, as well as the algebra $\mathcal{E}$ of finite disjoint unions of such rectangles; recall Exercise 4.13. On the other hand, the continuity properties (4.7) and (4.14) of the measure $\mathbb{P}_2$ from below and above, respectively (the latter needs the assumption $\mathbb{P}_2(\Omega_2) < \infty$), allow us to check that $\mathcal{M}$ is also a *monotone class* (Definition 4.1). Thus, from the Monotone Class Theorem 4.6 and Exercise 4.13, we obtain the inclusions $\mathcal{M} \supseteq m(\mathcal{E}) = \sigma(\mathcal{E}) = \sigma(\mathcal{R}) =: \mathcal{F}_1\otimes\mathcal{F}_2$.

• We can verify now that *the set-function* $\mathcal{F}_1\otimes\mathcal{F}_2 \ni E \to \mathbb{P}(E) \in [0,\infty)$ *of (5.16) is a finite measure which satisfies (5.13) on* $\mathcal{R}$. To see this latter property, just note that for $E = E_1\times E_2$ with $E_i \in \mathcal{F}_i$ $(i = 1, 2)$ we have: $E_{\omega_1} = E_2$ (resp., $\emptyset$) for $\omega_1 \in E_1$ (resp., for $\omega_1 \notin E_1$), thus

$$\mathbb{P}_2(E_{\omega_1}) = \mathbf{1}_{E_1}(\omega_1)\cdot\mathbb{P}_2(E_2)\,, \qquad \mathbb{P}(E) = \int_{\Omega_1}\mathbb{P}_2(E_{\omega_1})\,\mathrm{d}\,\mathbb{P}_1(\omega_1) = \mathbb{P}_1(E_1)\cdot\mathbb{P}_2(E_2)\,.$$

In particular, $\mathbb{P}(\Omega_1\times\Omega_2) = \mathbb{P}(\Omega_1)\cdot\mathbb{P}(\Omega_2) < \infty$.

To check countable additivity, take any sequence $\{E^{(n)}\}_{n\in\mathbb{N}} \subseteq \mathcal{F}_1 \otimes \mathcal{F}_2$ of disjoint sets, let $E = \bigcup_{n\in\mathbb{N}} E^{(n)}$, observe that $E_{\omega_1} = \bigcup_{n\in\mathbb{N}} \big(E^{(n)}\big)_{\omega_1}$ is again a disjoint union, and check

$$\mathbb{P}(E) \;=\; \int_{\Omega_1} \mathbb{P}_2(E_{\omega_1})\,\mathrm{d}\,\mathbb{P}_1(\omega_1) \;=\; \sum_{n\in\mathbb{N}} \int_{\Omega_1} \mathbb{P}_2\big((E^{(n)})_{\omega_1}\big)\,\mathrm{d}\,\mathbb{P}_1(\omega_1) \;=\; \sum_{n\in\mathbb{N}} \mathbb{P}(E^{(n)})\,.$$

For uniqueness, suppose that $\mathbb{P}_*$ is another measure on $\mathcal{F}_1 \otimes \mathcal{F}_2$, the product σ-algebra of $\Omega_1 \times \Omega_2$, which satisfies (5.13); then $\mathbb{P}_* \equiv \mathbb{P}$ on the $\pi-$system $\mathcal{R}$ of measurable rectangles, which generates $\mathcal{F}_1 \otimes \mathcal{F}_2 = \sigma(\mathcal{R})$. From Theorem 4.8, we deduce that $\mathbb{P}_* \equiv \mathbb{P}$ holds also on $\mathcal{F}_1 \otimes \mathcal{F}_2$.

Interchanging the rôles of the two indices in the preceding arguments, we see that $\widetilde{\mathbb{P}}(E) := \int_{\Omega_2} \mathbb{P}_1(E_{\omega_2})\,d\mathbb{P}_2(\omega)$ is a finite measure on $\mathcal{F}_1 \otimes \mathcal{F}_2$ that shares all the properties of $\mathbb{P}$; from uniqueness, $\widetilde{\mathbb{P}} \equiv \mathbb{P}$. This establishes (5.17).

• If $\mathbb{P}_1$, $\mathbb{P}_2$ are only $\sigma-$finite, we can write the product space $\Omega_1 \times \Omega_2$ as a countable, increasing union of disjoint rectangles $\Omega_1^{(n)} \times \Omega_2^{(n)}$ in $\mathcal{R}$, whose sides have finite measures. It suffices then to establish the result on *each* such rectangle, which we have already done; then pass to the limit, invoking the Monotone Convergence Theorem. □

Proof of Theorem 5.2 : If $X = \mathbf{1}_E$ for some $E \in \mathcal{F}_1 \otimes \mathcal{F}_2$, then

$$h_1(\omega_1) \equiv g_1(\omega_1) := \mathbb{P}_2\big(\omega_2 \in \Omega_2 \,:\, (\omega_1,\omega_2) \in E\big) = \mathbb{P}_2(E_{\omega_1})\,,$$

$$h_2(\omega_2) \equiv g_2(\omega_2) := \mathbb{P}_1\big(\omega_1 \in \Omega_1 \,:\, (\omega_1,\omega_2) \in E\big) = \mathbb{P}_1(E_{\omega_2})$$

and the result reduces to Theorem 5.1; so it holds for simple functions.

For general $X \in \mathbb{L}^+(\mathbb{P}_1 \otimes \mathbb{P}_2)$, let the sequence of simple functions $\{X^{(n)}\}_{n\in\mathbb{N}} \subset \mathbb{L}^+$ increase pointwise to X; recall Proposition 4.1. Then $h_i^{(n)} \uparrow h_i$ pointwise (in particular, h_i is measurable), and $\int_{\Omega_i} h_i\,\mathrm{d}\,\mathbb{P}_i = \lim_n \int_{\Omega_i} h_i^{(n)}\,\mathrm{d}\,\mathbb{P}_i = \lim_n \int_\Omega X^{(n)}\,d\mu = \int_\Omega X\,\mathrm{d}\,\mathbb{P}\,,\ i = 1,\,2\,.$ □

Proof of Theorem 5.3: If the function $X \in \mathbb{L}^1(\mathbb{P}_1 \otimes \mathbb{P}_2)$ is non-negative, we have from the TONELLI Theorem 5.2 that $h_i(\omega_i) < \infty$ (i.e., $X_{\omega_i} \in \mathbb{L}^1(\mathbb{P}_j)$) holds for $\mathbb{P}_i-$a.e. $\omega_i \in \Omega_i$; as well as $h_i \in \mathbb{L}^1(\mathbb{P}_i)$, for each $i = 1, 2$. In the general case, the result follows by applying the TONELLI Theorem 5.2 to each of X^+, X^- separately, then adding things up. □

Remark 5.1. It is straightforward to extend the above results to several dimensions: if $(\Omega_i, \mathcal{F}_i, \mathbb{P}_i)$, $i = 1, \cdots, d$ are $\sigma-$finite measure spaces, then we define as before the product σ-algebra $\mathcal{F} = \bigotimes_{i=1}^d \mathcal{F}_i$ as the smallest σ-algebra that contains the collection

$$\mathcal{R} \;:=\; \big\{E_1 \times \cdots \times E_d \;\big|\; E_1 \in \mathcal{F}_1\,, \cdots,\, E_d \in \mathcal{F}_n\big\}$$

of measurable rectangles, and construct on it a (unique) measure $\mathbb{P}$ with the property

$$\mathbb{P}\big(E_1 \times \cdots \times E_d\big) \;=\; \mathbb{P}_1(E_1) \cdots \mathbb{P}_d(E_d)\,, \qquad E_i \in \mathcal{F}_i\,,\ i = 1, \cdots, d\,. \tag{5.20}$$

This measure is denoted $\mathbb{P} = \otimes_{i=1}^d \mathbb{P}_i$ and is called the **product measure of** $\mathbb{P}_1, \cdots, \mathbb{P}_d$. Similarly, $\big(\prod_{i=1}^d \Omega_i, \bigotimes_{i=1}^d \mathcal{F}_i, \otimes_{i=1}^d \mathbb{P}_i\big)$ is then called the **product measure space** of $(\Omega_i, \mathcal{F}_i, \mathbb{P}_i)$, $i =$

$1, \cdots, d$. With the obvious modifications in notation, this measure has the properties set out in Theorems 5.1-5.3. In addition to the *commutativity property*

$$(\mathbb{P}_2 \otimes \mathbb{P}_1)(\mathcal{D}E) \;=\; (\mathbb{P}_1 \otimes \mathbb{P}_2)(E) \qquad \text{with} \quad \mathcal{D}E := \{\, (\omega_1, \omega_2) \,:\, (\omega_2, \omega_1) \in E \,\}\,, \quad \text{for } E \in \mathcal{E}$$

implied by Theorem 5.1, the product measure is *associative:* $\mathbb{P}_1 \otimes (\mathbb{P}_2 \otimes \mathbb{P}_3) \;=\; (\mathbb{P}_1 \otimes \mathbb{P}_2) \otimes \mathbb{P}_3$.

Remark 5.2. The TONELLI-FUBINI Theorems 6.2, 6.3 are most usefully invoked "in concatenation", in order to justify **inverting the order of integration** in double integrals of the form

$$\int_{\Omega_1} \int_{\Omega_2} X \,\mathrm{d}\mathbb{P}_1 \,\mathrm{d}\mathbb{P}_2 \;=\; \int_{\Omega} X \,\mathrm{d}(\mathbb{P}_1 \otimes \mathbb{P}_2)\,.$$

Typically, one verifies *first* that $\int_\Omega |X| \,\mathrm{d}(\mathbb{P}_1 \otimes \mathbb{P}_2)$ is finite, using TONELLI's theorem to evaluate this as a double integral – and *then* one applies FUBINI's theorem to conclude

$$\int_{\Omega_1} \left(\int_{\Omega_2} f \,\mathrm{d}\mathbb{P}_2 \right) \mathrm{d}\mathbb{P}_1 = \int_{\Omega_2} \left(\int_{\Omega_1} f \,\mathrm{d}\mathbb{P}_1 \right) \mathrm{d}\mathbb{P}_2\,.$$

As a (very good) rule of thumb, whenever you come across a double integral, *invert the order of integration!* Just do it; then worry about justifying what you did, using Theorems 5.3 and 5.2 as explained above.

Exercise 5.9. Layered Representation of Expectations: Suppose that ν is a measure on $\mathcal{B}([0,\infty))$ with $N(u) := \nu([0,u)) < \infty\,,\ \forall\ u > 0$, and that $X : \Omega \to [0,\infty)$ is a measurable function on the $\sigma-$finite measure space $(\Omega, \mathcal{F}, \mathbb{P})$. Show that

$$\int_\Omega N(X(\omega)) \,\mathrm{d}\mathbb{P}(\omega) \;=\; \int_{[0,\infty)} \mathbb{P}(X > u) \,\mathrm{d}\nu(u)\,. \tag{5.21}$$

In particular, if $\mathrm{d}\nu(u) = r u^{r-1} du$ for some $r > 0$, then

$$\boxed{\mathbb{E}(X^r) \;=\; \int_\Omega \big(X(\omega)\big)^r \,\mathrm{d}\mathbb{P}(\omega) \;=\; r \int_0^\infty u^{r-1}\, \mathbb{P}(X > u) \,\mathrm{d}u\,.} \tag{5.22}$$

The layered representation formula (5.22) allows us to compute painlessly the moments of the exponential distribution in Exercise 4.8: to wit, for $r > 0$ it gives

$$\mathbb{E}(X^r) \;=\; r \int_0^\infty u^{r-1}\, e^{-\lambda u} \,\mathrm{d}u \;=\; \frac{r}{\lambda^r} \int_0^\infty \xi^{\,r-1}\, e^{-\xi} \,\mathrm{d}\xi \;=\; \frac{r\,\Gamma(r)}{\lambda^r} \;=\; \frac{\Gamma(r+1)}{\lambda^r}\,.$$

Exercise 5.10. Boundedness of Linear Operators on $\mathbb{L}^p-$spaces: Let $(X, \mathcal{F}, \mu)$ and $(Y, \mathcal{G}, \nu)$ be σ-finite measure spaces, and $K : X \times Y \to \mathbb{R}$ an $(\mathcal{F} \otimes \mathcal{G})-$measurable function. Suppose that, for some $C \in [0,\infty)$, we have

(i) $\int_X |K(x,y)|\, d\mu(x) \le C\,,$ for $\nu-$a.e. $y \in Y$,

(ii) $\int_Y |K(x,y)|\, d\nu(y) \le C$, for $\mu-$a.e. $x \in X$.

Then for every $f \in \mathbb{L}^p(\nu)$, $1 \le p \le \infty$, the integral

$$(Tf)(x) := \int_Y K(x,y)\, f(y)\, \mathrm{d}\nu(y)$$

converges absolutely for $\mu-$a.e. $x \in X$; the function Tf is well-defined and in $\mathbb{L}^p(\mu)$; and we have the *Generalized* YOUNG*'s Inequality:* $||Tf||_p \le C\, ||f||_p$.

Exercise 5.11. Convolution and the FOURIER Transform: For any two measurable functions $f : \mathbb{R}^d \to \mathbb{R}$ and $g : \mathbb{R}^d \to \mathbb{R}$, the *convolution* $f \star g$ of f and g is the function defined by

$$(f \star g)(x) := \int_{\mathbb{R}^d} f(x-y)\, g(y)\, \mathrm{d}y = \int_{\mathbb{R}^d} g(x-y)\, f(y)\, \mathrm{d}y \tag{5.23}$$

for all $x \in \mathbb{R}^d$ such that the integral on the right-hand-side is well-defined and finite. For instance, if $f \in \mathbb{L}^p(\mathbb{R}^d)$ and $q \in \mathbb{L}^q(\mathbb{R}^d)$ with $p \ge 1$, $(1/p) + (1/q) = 1$, then the Hölder inequality guarantees that $(f * g)(x)$ is well-defined and finite for every $x \in \mathbb{R}^d$.
(i) Assuming that all integrals in question exist, show that

$$f \star g = g \star f\,, \quad \big(f \star g\big) \star h = f \star \big(g \star h\big)\,.$$

(ii) Show that, for every $g \in \mathbb{L}^1(\mathbb{R}^d)$ and $f \in \mathbb{L}^p(\mathbb{R}^d)$ for some $1 \le p \le \infty$, the convolution $(f \star g)(x)$ of (5.23) is well-defined for $\lambda-$a.e. $x \in \mathbb{R}^d$, and satisfies *Young's inequality*

$$||f \star g||_p \le ||g||_1\, ||f||_p\,.$$

(iii) With $i = \sqrt{-1}$, the *Fourier Transform* of $f \in \mathbb{L}^1(\mathbb{R}^d)$ is the function $\widehat{f} : \mathbb{R}^d \to \mathbb{C}$ defined by

$$\widehat{f}(\xi) = \int_{\mathbb{R}} e^{\,i\langle \xi, x\rangle}\, f(x)\, \mathrm{d}x\,, \quad \xi \in \mathbb{R}^d\,. \tag{5.24}$$

Show that $\widehat{f}(\cdot)$ is uniformly continuous, and uniformly bounded: $\sup_{\xi \in \mathbb{R}^d} |\widehat{f}(\xi)| \le ||f||_1 < \infty$. Show also that *the* FOURIER *transform of the convolution is the product of the* FOURIER *transforms*, in the sense that with $h := f \star g$ we have

$$\widehat{h} = \widehat{f} \cdot \widehat{g} \quad \text{for every} \quad f \in \mathbb{L}^1(\mathbb{R}^d)\,,\ g \in \mathbb{L}^1(\mathbb{R}^d)\,.$$

5.5 LEBESGUE Decomposition and RADON-NIKODÝM

Suppose that $\mu,\, \nu$ are two measures on the *same* measurable space $(\Omega, \mathcal{F})$. We say that ν *is absolutely continuous with respect to* μ, and write $\nu < \mu$, if $A \in \mathcal{F}$ and $\mu(A) = 0$ imply $\nu(A) = 0$. For example, Exercise 4.19 shows that this is the case when

$$\nu(A) \,=\, \int_A h \,\mathrm{d}\mu\,, \quad \forall\ A \in \mathcal{F}\,, \quad \text{for some } \ h \in \mathbb{L}^1(\mu) \cap \mathbb{L}^0_+\,. \tag{5.25}$$

We shall see below that this example is not so special: to wit, if μ is $\sigma-$finite and ν is finite, then $\nu < \mu$ implies (5.25). This is the content of the celebrated RADON-NIKODÝM Theorem 5.5.

We say that $\mu,\ \nu$ are *equivalent* (and write $\mu \sim \nu$), if they are mutually absolutely continuous, that is $\nu < \mu$ and $\mu < \nu$. Finally, we say that $\mu,\ \nu$ are *singular* (and write $\mu \perp \nu$), if there exists a measurable set $A \in \mathcal{F}$ such that $\mu(A) = \nu(A^c) = 0$.

Theorem 5.4. LEBESGUE Decomposition: *Let $(\Omega, \mathcal{F})$ be a measurable space, and that $\mu,\ \nu$ are two σ-finite measures on it. Then there exist measures $\nu_{ac}\,,\ \nu_s$ with*

$$\nu = \nu_{ac} + \nu_s\,; \quad \nu_{ac} < \mu\,, \quad \nu_s \perp \mu\,,$$

and this decomposition is unique.

For instance, let $\lambda\big|_{[a,b]}$ denote Lebesgue measure on the interval $[a, b]$, and take $\mu = \lambda\big|_{[0,2]}$, $\nu = \lambda\big|_{[1,3]}$. Then $\nu_{ac} = \lambda\big|_{[1,2]}$ and $\nu_s = \lambda\big|_{(2,3]}$.

Theorem 5.5. RADON-NIKODÝM: *Let the two measures $\mu,\ \nu$ on the measurable space $(\Omega, \mathcal{F})$ be σ-finite and finite, respectively, with $\nu < \mu$. Then there exists a unique (up to $\mu-$a.e. equivalence) function $h \in \mathbb{L}^1(\mu) \cap \mathbb{L}^0_+$ such that, as in (5.25), we have*

$$\nu(A) = \int_A h \,\mathrm{d}\mu\,, \quad \forall\, A \in \mathcal{F}\,.$$

The function $h : \Omega \to [0, \infty)$ of (5.25) is called the RADON-NIKODÝM *derivative* of ν with respect to μ, and is denoted $h = (\mathrm{d}\nu/\mathrm{d}\mu)$, in much the same manner as (5.25) is commonly expressed in the compact form $\mathrm{d}\nu = h\,\mathrm{d}\mu$.

This notation suggests correct conclusions. For instance, if $\nu < \mu$ and $f \in \mathbb{L}^1(\nu)$, then $\int_\Omega f\ (\mathrm{d}\nu/\mathrm{d}\mu)\,\mathrm{d}\mu = \int_\Omega f\,\mathrm{d}\nu$; and if $\rho < \nu < \mu$ are finite measures, then we have

$$(\mathrm{d}\rho/\mathrm{d}\mu) = (\mathrm{d}\rho/\mathrm{d}\nu) \cdot (\mathrm{d}\nu/\mathrm{d}\mu)\ , \quad \mu - \text{a.e.}$$

We shall prove Theorems 5.4 and 5.5 in chapter 16 (Appendix).

5.6 Completeness of the LEBESGUE Spaces $\mathbb{L}^p$, $1 \leq p \leq \infty$

The following result shows that the normed linear spaces $\mathbb{L}^p$ of this section are BANACH spaces (that is, *complete* in the topologies induced by the norms of (5.6), (5.11) for $1 \leq p \leq \infty$. A brief overview of the definitions and basic properties regarding BANACH and HILBERT spaces appears in chapter 16 (Appendix).

Theorem 5.6. All spaces $\mathbb{L}^p$, $1 \leq p \leq \infty$ are complete. *In other words: For any* CAUCHY sequence $\{X_n\}_{n\in\mathbb{N}}$ *in* $\mathbb{L}^p$, *to wit, with the property that for every* $\varepsilon > 0$ *there is an integer* N_ε *so that*

$$||X_n - X_m||_p \leq \varepsilon \quad \textit{holds for any } n \geq N_\varepsilon\,,\ m \geq N_\varepsilon\,, \tag{5.26}$$

there exists a unique $X \in \mathbb{L}^p$ *such that* $||X_n - X||_p \to 0$ *as* $n \to \infty$.

Furthermore, there exists a subsequence $\{X_{n_k}\}_{k\in\mathbb{N}} \subseteq \{X_n\}_{n\in\mathbb{N}}$, *as well as a function* $F : \Omega \to [0,\infty)$ *in* $\mathbb{L}^p$, *such that for* $\mathbb{P}-$*a.e.* $\omega \in \Omega$ *we have:*

$$|X_{n_k}(\omega)| \leq F(\omega)\,,\ \forall\, k \in \mathbb{N} \quad \textit{and} \quad \lim_{k\to\infty} X_{n_k}(\omega) = X(\omega)\,.$$

The argument involves a couple of ideas that are often used to great advantage in Analysis and in Probability. The first idea is that

- subsequences that converge "sufficiently fast" in $\mathbb{L}^p$ must converge also $\mathbb{P}-$a.e.;

whereas the second idea is that

- it is enough first to show $\mathbb{L}^p-$convergence for *some* subsequence, then use a "sandwich argument" to argue $\mathbb{L}^p-$convergence for *the entire* sequence.

Proof: To see how these ideas work in our present context, observe that the CAUCHY property (5.26) allows us to choose a subsequence $\{X_{n_k}\}$ with $||X_{n_{k+1}} - X_{n_k}||_p \leq 2^{-k}$, $\forall\ k \in \mathbb{N}$. The sequence of positive functions $\{F_k\}_{k\in\mathbb{N}}$ defined by

$$F_1(\omega) := |X_{n_1}(\omega)|\,, \quad F_{k+1}(\omega) := |X_{n_1}(\omega)| + \sum_{j=1}^{k} |X_{n_{j+1}}(\omega) - X_{n_j}(\omega)| \quad \text{for } k \in \mathbb{N}$$

satisfies $||F_k||_p \leq ||X_{n_1}||_p + \sum_{j=1}^{k-1} 2^{-j} \leq ||X_{n_1}||_p + 1$ from the triangle inequality, and increases $\mathbb{P}-$a.e. to a function F; then FATOU's Lemma guarantees that F is in $\mathbb{L}^p$, hence also $\mathbb{P}-$a.e. finite: $\mathbb{E}(F^p) \leq \liminf_{k\to\infty} \mathbb{E}((F_k)^p) \leq \big(1 + ||X_{n_1}||_p\big)^p < \infty$.

As a result, for $\mathbb{P}-$a.e. $\omega \in \Omega$, the sequence of real numbers

$$X_{n_{k+1}}(\omega) = X_{n_1}(\omega) + \sum_{j=1}^{k} \big[X_{n_{j+1}}(\omega) - X_{n_j}(\omega)\big]\,, \qquad k \in \mathbb{N}$$

converges absolutely to some real number

$$X(\omega) := \lim_{k\to\infty} X_{n_k}(\omega) = X_{n_1}(\omega) + \sum_{j\in\mathbb{N}} \big[X_{n_{j+1}}(\omega) - X_{n_j}(\omega)\big]\,.$$

Because $|X_{n_k}(\omega)| \leq F(\omega)$ and $F \in \mathbb{L}^p$, we deduce from the Dominated Convergence Theorem that $X \in \mathbb{L}^p$ and

$$||X_{n_k} - X||_p \longrightarrow 0\,, \quad \text{as} \quad k \to \infty\,, \tag{5.27}$$

since $|X_{n_k} - X| \le F + |X| \in \mathbb{L}^p$.

Now let us argue that the *entire sequence* $\{X_n\}_{n\in\mathbb{N}}$ *must converge in* $\mathbb{L}^p$ *to this function* $X \in \mathbb{L}^p$. Indeed, for any $\varepsilon > 0$ we can choose on account of (5.27) an integer $K_\varepsilon \in \mathbb{N}$ large enough, so that $||X_{n_k} - X||_p \le \varepsilon/2$ holds for all $k \ge K_\varepsilon$. On the other hand, from (5.26) we can choose $N_\varepsilon \in \mathbb{N}$ large enough, so that $||X_n - X_{n_k}||_p \le \varepsilon/2$ holds for all $n \ge N_\varepsilon$ and for all $k \ge K_\varepsilon$. The triangle inequality now implies

$$||X_n - X||_p \le ||X_n - X_{n_k}||_p + ||X_{n_k} - X||_p \le \varepsilon\,, \quad \forall\, n \ge N_\varepsilon\,,$$

which shows that the entire sequence $\{X_n\}_{n\in\mathbb{N}}$ converges in $\mathbb{L}^p$ to the function X. □

The HÖLDER inequality places the BANACH spaces of this theorem in a formal *duality,* with $\mathbb{L}^q$ the dual of the space $\mathbb{L}^p$ for $1 \le p \le \infty$ when $(1/p) + (1/q) = 1$.

Clearly, the space $\mathbb{L}^2$ of square-integrable functions is *self-dual* in this sense; it is also a *Hilbert space* with inner product $\langle f, g\rangle = \int_\Omega f g\, d\mu$, as discussed in Chapter 16 (Appendix).

Exercise 5.12. If $1 \le p < \infty$, the simple functions of the form $f = \sum_{n=1}^N \alpha_n \mathbf{1}_{E_n}$ with $\alpha_n \in \mathbb{R}$ and $\mathbb{P}(E_n) < \infty$ for $n = 1, \ldots, N$, $N \in \mathbb{N}$ are dense in $\mathbb{L}^p$.

Exercise 5.13. Justifying the notation $||X||_\infty$ for the essential least-upper-bound. Show that if $X \in \mathbb{L}^r \cap \mathbb{L}^\infty$ for some $1 \le r < \infty$, then $X \in \mathbb{L}^p$ for any $p \in [r, \infty]$ and we have

$$||X||_\infty = \lim_{p\to\infty} ||X||_p\,.$$

6 Constructing Measure Spaces

Let us return now to the question broached briefly at the end of section 2.4. Suppose we are given a nondecreasing, right continuous function $F : \mathbb{R} \to \mathbb{R}$. *How do we construct on the collection* $\mathcal{B}(\mathbb{R})$ *of* BOREL *subsets of the real line a measure* μ_F, *to be called* LEBESGUE-STIELTJES *measure associated with* $F(\cdot)$, *so that*

$$\mu_F(I) = F(b) - F(a) \quad \textit{holds for every interval } I = (a, b]\,? \tag{6.1}$$

As we saw in Exercise 4.14, it is necessary to "lower our sights" and accept the fact that we have to do this job on a $\sigma-$algebra of subsets of the real line that is *strictly smaller* than its power set $\mathcal{P}(\mathbb{R})$. The $\sigma-$algebra $\mathcal{B}(\mathbb{R})$ of BOREL subsets is then definitely the next best thing around.

It is fairly clear how to start. We consider the collection $\mathcal{G}$ of intervals of the type $(a, b]$ with $-\infty \le a \le b < \infty$; recall from the discussion following Definition 4.3 that $\mathcal{G}$ is the prototype of what was called there "elementary family" [35]. Indeed, the intersection of two intervals from $\mathcal{G}$ is an interval in $\mathcal{G}$; whereas the complement of an interval from $\mathcal{G}$ is either an interval in $\mathcal{G}$ or the union of two disjoint intervals from $\mathcal{G}$. In accordance with (6.1), we define on $\mathcal{G}$ the set function $\ell_F : \mathcal{G} \to [0, \infty)$ by the recipe

$$\ell_F\big((a, b]\big) \;:=\; F(b) - F(a)$$

We regard this as a type of "proto-measure" associated with the function $F(\cdot)$.

We look next at the collection $\mathcal{E}$ of all finite unions $\bigcup_{j=1}^n (a_j, b_j]$ of disjoint intervals from $\mathcal{G}$, say with $\infty > b_1 > a_1 \ge \cdots \ge b_{n-1} > a_{n-1} \ge b_n > a_n \ge -\infty$. This collection is easily seen to be an algebra (cf. Exercise 4.13), and we define on it the set function $\nu : \mathcal{E} \to [0, \infty)$ via

$$\nu_F\Big(\bigcup_{j=1}^n (a_j, b_j]\Big) \;:=\; \sum_{j=1}^n \ell_F\big((a_j, b_j]\big) \;=\; \sum_{j=1}^n [\,F(b_j) - F(a_j)\,]\,, \quad \big\{(a_j, b_j]\big\}_{j=1}^n \subseteq \mathcal{G}\,. \tag{6.2}$$

Since $F(\cdot)$ is nondecreasing this set function $\nu \equiv \nu_F$ is nonnegative and finitely additive, and agrees with ℓ_F on $\mathcal{G}$. It is also seen easily that ν is well-defined by (6.2), in the sense that $\nu(E)$ *assigned by (6.2) is independent of the way in which a given set* $E \in \mathcal{E}$ *is partitioned into a finite union of disjoint intervals.*

Indeed, if $(a_1, b_1], \cdots, (a_n, b_n]$ are disjoint intervals and $\bigcup_{j=1}^n (a_j, b_j] = (a, b]$, then we have $b = b_1 > a_1 = b_2 > \cdots > a_{n-1} = b_n > a_n = a$ (by relabeling some indices if necessary) and $\sum_{j=1}^n [F(b_n) - F(a_n)] = F(b) - F(a)$. More generally, let $\{I_i\}_{1 \le i \le m}$ and $\{J_j\}_{1 \le j \le n}$ be finite collections of disjoint intervals in $\mathcal{G}$ such that $\bigcup_{i=1}^m I_i = \bigcup_{j=1}^n J_j$; then this same reasoning gives

$$\sum_{i=1}^m \nu(I_i) = \sum_{i=1}^m \sum_{j=1}^n \nu(I_i \cap J_j) = \sum_{j=1}^n \nu(J_j)\,,$$

so ν is indeed well defined.

[35] Reminder (Definition 4.3): An *elementary family* is a collection of sets that contains the empty set, is closed under finite intersections, and has the property that the complement of every set in the collection can be written as the union of finitely many, pairwise disjoint sets in the collection.

• Next, we should like to extend the set function $\nu : \mathcal{E} \to [0,\infty)$ of (10.1) to a measure $\mu \equiv \mu_F$ on the $\sigma-$algebra $\mathcal{B}(\mathbb{R}) = \sigma(\mathcal{E}) = \sigma(\mathcal{G})$ of BOREL subsets of the real line. In order to do this we shall need the following fundamental result, whose proof we defer to Chapter 14 (Appendix).

Theorem 6.1. CARATHÉODORY-HAHN Extension: *Suppose we are given an algebra $\mathcal{E}$ of subsets of a nonempty set Ω, and on it a $\sigma-$finite set function $\nu : \mathcal{E} \to [0,\infty)$ which is a* pre-measure: *to wit, satisfies $\nu(\emptyset) = 0$ and the property $\nu\big(\bigcup_{n\in\mathbb{N}} E_n\big) = \sum_{n\in\mathbb{N}} \nu(E_n)$ for any sequence $\{E_n\}_{n\in\mathbb{N}} \subseteq \mathcal{E}$ of pairwise disjoint sets with $\bigcup_{n\in\mathbb{N}} E_n \in \mathcal{E}$.*

There exists then a unique *measure $\mu : \mathcal{F} \to [0,\infty)$ on the $\sigma-$algebra $\mathcal{F} := \sigma(\mathcal{E})$ generated by $\mathcal{E}$, with the property $\mu \equiv \nu$ on $\mathcal{E}$. In other words, this μ is the unique extension of ν to a measure on $\mathcal{F}$. Furthermore, this measure has the "outer approximation" property*

$$\mu(A) \;=\; \inf\left\{ \sum_{n\in\mathbb{N}} \nu(E_n) \;\middle|\; \{E_n\}_{n\in\mathbb{N}} \subseteq \mathcal{E}\,,\; A \subseteq \bigcup_{n\in\mathbb{N}} E_n \right\}, \qquad A \in \mathcal{F}\,. \tag{6.3}$$

Let us apply this result to our situation at hand. We need to verify that ν_F, as defined in (6.2), *is a pre-measure on $\mathcal{E}$*; in other words, that for every collection of disjoint intervals $\{(a_n, b_n]\}_{n\in\mathbb{N}} \subseteq \mathcal{G}$ with $\bigcup_{n\in\mathbb{N}}(a_n, b_n] \in \mathcal{E}$, we have the countable additivity property

$$\nu_F\Big(\bigcup_{n\in\mathbb{N}}(a_n, b_n]\Big) \;=\; \sum_{n\in\mathbb{N}} \ell_F\big((a_n, b_n]\big)\,.$$

But we are assuming that the union $E = \bigcup_{n\in\mathbb{N}}(a_n, b_n]$ belongs to the algebra $\mathcal{E}$, so this set E is also the union of *finitely many* disjoint intervals from $\mathcal{G}$. Therefore, we may partition the collection $\{(a_n, b_n]\}_{n\in\mathbb{N}}$ into finitely many sub-collections, in such a way that the union of the intervals in each of these sub-collections is a *single* interval in $\mathcal{G}$.

Thus, it is enough to establish the following.

Proposition 6.1. *Whenever an interval $(a, b]$ can be written as a countable disjoint union $(a, b] = \bigcup_{n\in\mathbb{N}}(a_n, b_n]$, we have*

$$\sum_{n\in\mathbb{N}} \ell_F((a_n, b_n]) = \ell_F((a, b]).$$

Proof: It is straightforward to check that, whenever $(a, b] = \bigcup_{n=1}^{N}(a_n, b_n]$ is a *finite* disjoint union, the additivity property $\ell_F((a, b] = \sum_{n=1}^{N} \ell_F((a_n, b_n])$ holds. Indeed, one establishes separately the implications

$$\bigcup_{n=1}^{N}(a_n, b_n] \subseteq (a, b] \quad\Longrightarrow\quad \sum_{n=1}^{N} \ell_F((a_n, b_n]) \;\le\; \ell_F((a, b])\,, \tag{6.4}$$

$$(a, b] \subseteq \bigcup_{n=1}^{N}(a_n, b_n] \quad\Longrightarrow\quad \ell_F((a, b]) \;\le\; \sum_{n=1}^{N} \ell_F((a_n, b_n]) \tag{6.5}$$

(the second does not even need the intervals to be disjoint), and the finite additivity follows.

For an infinite collection $\{(a_n, b_n]\}_{n\in\mathbb{N}}$ of disjoint intervals in $\mathcal{G}$ with $(a, b\,] = \bigcup_{n\in\mathbb{N}}(a_n, b_n]$, we have clearly $\bigcup_{n=1}^{N}(a_n, b_n] \subseteq (a, b\,]$ for every $N \in \mathbb{N}$. We let now $N \to \infty$ in the first of the above implications, namely (6.4), and obtain $\sum_{n\in\mathbb{N}} \ell_F((a_n, b_n]) \leq \ell_F((a, b\,])$.

• To obtain also the reverse inequality and finish the proof, it seems like a very obvious step to pass from a finite to an infinite collection. Yet the fact that the equality $(a, b\,] = \bigcup_{n\in\mathbb{N}}(a_n, b_n]$ holds, does not necessarily give $(a, b\,] \subseteq \bigcup_{n=1}^{N}(a_n, b_n]$ as in (6.5), for any given $N \in \mathbb{N}$. In order to be able to make such a step, one needs (some version of) the HEINE-BOREL theorem; [36] see, for instance, KOLMOGOROV & FOMIN (1970), page 92.

One way to do this is to select, for any given $\varepsilon \in (0, b-a)$ and $n \in \mathbb{N}$, a real number $b'_n > b_n$ with $F(b'_n) - F(b_n) < \varepsilon 2^{-n}$ (the right-continuity of $F(\cdot)$ is needed here in a crucial manner). Then $[a+\varepsilon, b] \subseteq \bigcup_{n\in\mathbb{N}}(a_n, b'_n)$, so by the HEINE-BOREL theorem there exists a *finite* collection $(a_{n_1}, b'_{n_1}), \cdots, (a_{n_K}, b'_{n_K})$ of open intervals which covers $[a+\varepsilon, b\,]$, in the sense that

$$(a+\varepsilon, b] \subset [a+\varepsilon, b] \subseteq \bigcup_{k=1}^{K}(a_{n_k}, b'_{n_k}) \subseteq \bigcup_{k=1}^{K}(a_{n_k}, b'_{n_k}\,];$$

and from (6.5) of our previous discussion, we get

$$F(b) - F(a+\varepsilon) \;=\; \ell_F\big((a+\varepsilon, b\,]\big) \;\leq\; \sum_{k=1}^{K} \ell_F\big((a_{n_k}, b'_{n_k}\,]\big)$$

$$\leq \sum_{k=1}^{K} \big[\, \ell_F\big((a_{n_k}, b_{n_k}\,]\big) + \varepsilon 2^{-n_k} \,\big] \;\leq\; \sum_{n\in\mathbb{N}} \ell_F\big((a_n, b_n\,]\big) + \varepsilon\,.$$

Letting $\varepsilon \downarrow 0$ and using again the right-continuity of $F(\cdot)$, we arrive at the desired inequality $F(b) - F(a) \;=\; \ell_F\big((a, b\,]\big) \;\leq\; \sum_{n\in\mathbb{N}} \ell_F\big((a_n, b_n\,]\big)$. □

Theorem 6.1 asserts that this construction leads to a unique measure $\mu \equiv \mu_F$ on $\mathcal{B}(\mathbb{R}) = \sigma(\mathcal{E})$ with the property (6.1), as well as

$$\mu_F(E) \;=\; \inf\left\{ \sum_{n\in\mathbb{N}} \big[\, F(b_n) - F(a_n) \,\big] \;\Big|\; E \subseteq \bigcup_{n\in\mathbb{N}} (a_n, b_n] \right\}, \qquad E \in \mathcal{B}(\mathbb{R})\,. \tag{6.6}$$

This is the so-called "LEBESGUE-STIELTJES measure" induced on the BOREL sets of the real line by the nondecreasing, right-continuous function $F : \mathbb{R} \to \mathbb{R}$. Integrals with respect to this measure are denoted as

$$\int_{(a,b]} f\, d\mu_F \;\equiv\; \int_a^b f(x)\, dF(x)\,, \qquad \text{for } -\infty \leq a < b < \infty\,. \tag{6.7}$$

[36] To the effect that *every cover of a closed interval of the form* $[a, b]$ *with* $-\infty < a < b < \infty$ *by a collection of open intervals, has a finite subcover.* This is the property of "compactness".

For the choice $F(x) \equiv x$, the resulting $\lambda \equiv \mu_F$ is the LEBESGUE *measure on the real line*. This measure is invariant under translations and dilations, in the sense

$$\lambda(E+s) \,=\, \lambda(E)\,, \quad \lambda(\,r\,E\,) \,=\, r\,\lambda(E)$$

for every $E \in \mathcal{B}(\mathbb{R})\,,\ s \in \mathbb{R}\,,\ r > 0\,,$ in addition of course to the property $\lambda((a,b]) = b-a\,.$

Discussion: Every singleton $\{x\}$ with $x \in \mathbb{R}$ has LEBESGUE measure zero; thus the same is true for every countable set. Enumerate as $\mathbf{Q}_1 = \{\varrho_n\}_{n\in\mathbb{N}} \subset [0,1]$ all the rational numbers of the unit interval, observe that $\lambda(\mathbf{Q}_1) = 0$ from the above discussion, and for any given $\varepsilon > 0$ set $I_n = \big(\varrho_n - \varepsilon 2^{-(n+1)}, \varrho_n + \varepsilon 2^{-(n+1)}\big)\,.$ Then the set $G := (0,1) \cap (\cup_{n\in\mathbb{N}} I_n)$ is dense (i.e., topologically "large") in the unit interval, as its closure $\overline{G}$ is the entire interval [0,1]. But is measure-theoretically "minuscule": its LEBESGUE measure is $\lambda(G) \le \sum_{n\in\mathbb{N}} \varepsilon 2^{-n} \le \varepsilon\,.$

6.1 Measures on Euclidean Spaces

Let us try to examine how this theory can be extended to higher-dimensional Euclidean spaces $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d)\,)$ with $d \ge 2$. Suppose that μ is a $\sigma-$finite measure on this space, and consider the function $F(\mathbf{x}) := \mu((-\infty, \mathbf{x}]),\ \mathbf{x} \in \mathbb{R}^d$ with the notation

$$\mathbf{a} \le \mathbf{b} \ \Leftrightarrow_{\text{def}}\ a_i \le b_i\,,\ \forall\, i = 1, \cdots, d \quad \text{and} \quad (a,b] := \{\omega \in \mathbb{R}^d \,|\, a_i < \omega_i \le b_i,\ \forall\, i = 1, \cdots, d\}$$

for vectors $\omega = (\omega_1, \cdots, \omega_d) \in \mathbb{R}^d$. The so-defined function $F : \mathbb{R}^d \to [0,\infty)$ is called *cumulative distribution function of* μ.

Setting $\Delta_{b_i - a_i}\, g(a_1, \cdots, a_{i-1}, a_i, a_{i+1}, \cdots, a_d) := g(a_1, \cdots, a_{i-1}, b_i, a_{i+1}, \cdots, a_d) - g(a_1, \cdots, a_{i-1}, a_i, a_{i+1}, \cdots, a_d)$ and $\Delta_{\mathbf{b}-\mathbf{a}}\, g(\mathbf{a}) := \Delta_{b_1-a_1} \cdots \Delta_{b_d - a_d}\, g(a_1, \cdots \cdots, a_d)\,,$ we observe that this function satisfies

$$\lim_{\substack{\mathbf{b}\to\mathbf{a} \\ \mathbf{a}\le\mathbf{b}}} F(\mathbf{b}) \,=\, F(\mathbf{a})\,, \quad \forall\ \mathbf{a} \in \mathbb{R}^d\,, \tag{6.8}$$

as well as $\Delta_{\mathbf{b}-\mathbf{a}}\, F(\mathbf{a}) = \mu((\mathbf{a},\mathbf{b}]) \ge 0\,,\ \ \forall\, \mathbf{a} \le \mathbf{b}$ in $\mathbb{R}^d\,.$

For instance, in the case $d = 2$ the left-hand side of this last expression is

$$\begin{aligned} \Delta_{b_1-a_1}\,(\Delta_{b_2-a_2}\, F(a_1,a_2)) \,&=\, \Delta_{b_2-a_2}\, F(b_1,a_2) - \Delta_{b_2-a_2}\, F(a_1,a_2) \\ &= [F(b_1,b_2) - F(b_1,a_2)] - [F(a_1,b_2) - F(a_1,a_2)] = \mu((\mathbf{a},\mathbf{b}])\,. \end{aligned}$$

We have the following generalization of the notions in Definition 10.1.

Definition 6.1. Distribution Function on $\mathbb{R}^d$: A BOREL-measurable function $F : \mathbb{R}^d \to [0,\infty)$ that satisfies the right-continuity property (6.8), as well as

$$\Delta_{\mathbf{b}-\mathbf{a}}\, F(\mathbf{a}) \ge 0\,, \quad \forall\, \mathbf{a} \le \mathbf{b} \quad \text{in } \mathbb{R}^d\,, \tag{6.9}$$

is called a distribution function on $\mathbb{R}^d$. A distribution function that satisfies also

$$\lim_{x_1\uparrow\infty,\,\cdots,\,x_d\uparrow\infty} F(x_1,\cdots,x_d) = 1 \quad \text{and} \quad \lim_{x_i\downarrow-\infty} F(x_1,\cdots,x_d) = 0\,,\ \forall\ i = 1,\cdots,d$$

is called a *probability distribution function* $\mathbb{R}^d\,.$

For $d = 1$, the distribution functions are the right-continuous, increasing functions. For $d \geq 2$, a function $F : \mathbb{R}^d \to [0, \infty)$ can be increasing and right-continuous in each of its variables separately, but fail to be a distribution function. Consider, for instance, the function $F(x_1, x_2) = \mathbf{1}_{\{x_1+x_2 \geq 0\}}$ on $\mathbb{R}^2$; with $\mathbf{a} = (-1, -1)$ and $\mathbf{b} = (2, 2)$ we have $\Delta_{\mathbf{b}-\mathbf{a}} F(\mathbf{a}) = F(2, 2) - F(2, -1) - F(-1, 2) + F(-1, -1) = -1$.

Definition 6.2. Product Distribution Function: If $F_1, \cdots, F_d$ are distribution functions on $\mathbb{R}$, then

$$F(\mathbf{x}) \equiv F(x_1, \cdots, x_d) := F_1(x_1) \cdots F_d(x_d), \quad \mathbf{x} \in \mathbb{R}^d \tag{6.10}$$

is a distribution function on $\mathbb{R}^d$, called the *product distribution function of* $F_1, \cdots, F_d$. If each $F_1, \cdots, F_d$ is a probability distribution function, then so is the product $F : \mathbb{R}^d \to [0, 1]$ of (6.10).

We can follow the now familiar procedure. For a given distribution function $F : \mathbb{R}^d \to \mathbb{R}$, we define $\nu((\mathbf{a}, \mathbf{b}]) := \Delta_{\mathbf{b}-\mathbf{a}} F(\mathbf{a})$ on the class $\mathcal{G}$ of rectangles of the form $(\mathbf{a}, \mathbf{b}]$. We allow (some of) the coördinates of $\mathbf{b}$ to become $+\infty$, and (some of) the coördinates of $\mathbf{a}$ to become $-\infty$, by passing to the appropriate limits; in such cases, we are effectively replacing $(\mathbf{a}, \mathbf{b}]$ by $(\mathbf{a}, \mathbf{b}] \cap \mathbb{R}^d$, whenever the former appears; we also allow $\mathbf{a} = \mathbf{b}$, that is, for the rectangle to become the empty set.

Consider then the algebra $\mathcal{E}$ consisting of finite disjoint unions of such rectangles, and extend ν to $\mathcal{E}$ by the recipe of (6.2). It can be seen, thanks to the conditions of Definition 6.1, that ν is a pre-measure on $\mathcal{E}$, so that Theorem 6.1 can be invoked to ensure that ν has a unique extension μ_F to the $\sigma-$algebra of BOREL sets $\mathcal{B}(\mathbb{R}^d) = \sigma(\mathcal{E})$ of $\mathbb{R}^d$. This is the LEBESGUE-STIELTJES *measure induced by* F **on** $\mathcal{B}(\mathbb{R}^d)$. Furthermore, if F is a probability distribution function, then μ_F is a probability measure. By analogy with (6.7), integrals with respect to this measure are denoted as

$$\int_{(\mathbf{a},\mathbf{b}]} f \, \mathrm{d}\mu_F \equiv \int_{a_1}^{b_1} \cdots \int_{a_d}^{b_d} f(x_1, \cdots, x_d) \, \mathrm{d}F(x_1, \cdots, x_d), \tag{6.11}$$

for $-\infty \leq a_i < b_i < \infty$, $i = 1, \cdots, d$.

Example 6.1. LEBESGUE Measure on $\mathcal{B}(\mathbb{R}^d)$*:* The measure λ, induced by the product distribution function $F(x) = x_1 \cdots x_d$ on $\mathcal{B}(\mathbb{R}^d)$, assigns to each rectangle its volume

$$\lambda((a, b]) = \prod_{i=1}^{d} (b_i - a_i), \qquad (a, b] \in \mathcal{G}.$$

This is called LEBESGUE *measure on* $\mathcal{B}(\mathbb{R}^d)$ and is translation-, reflection- and rotation-invariant.

The theory we have just developed allows us to construct, for any given distribution function $F(\cdot)$, a random vector that has this given $F(\cdot)$ as its probability distribution function. In Proposition 6.3 below, a special construction is presented for the one-dimensional case $d = 1$ which assumes only the existence of LEBESGUE measure on $\mathcal{B}(\mathbb{R})$.

Proposition 6.2. *For any given probability distribution function* $F : \mathbb{R}^d \to [0, 1]$ *there exists a probability space* $(\Omega, \mathcal{F}, \mathbb{P})$ *and a random vector* $\mathfrak{X} : \Omega \to \mathbb{R}^d$*, such that* $F_{\mathfrak{X}}(\cdot) \equiv F(\cdot)$.

Proof: An obvious choice is to take $\mathfrak{X}(\omega) = \omega$, the identity mapping on the space $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \mu_F)$ with μ_F the LEBESGUE-STIELTJES measure of section 6.1 corresponding to the given function $F(\cdot)$, and note $F_{\mathfrak{X}}(\mathbf{x}) = \mathbb{P}(\mathbf{X} \in (-\infty, \mathbf{x}]) = \mu_F((-\infty, \mathbf{x}]) = F(\mathbf{x}), \ \forall \, \mathbf{x} \in \mathbb{R}^d$.

6.2 SKOROHOD Construction

Suppose we are given a distribution function $F : \mathbb{R} \to \mathbb{R}$ (nondecreasing, right-continuous), and are asked to construct a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and on it a random variable X with the given $F(\cdot)$ as its probability distribution function, that is, $F_X(\cdot) \equiv F(\cdot)$.

Now, once LEBESGUE-STIELTJES measure has been constructed on the real line, this is just a very special case of Proposition 6.2 for $d = 1$. Let us present, however, a beautiful construction, due to A.V. SKOROHOD, which assumes only the existence of LEBESGUE measure on the unit interval $\big([0,1], \mathcal{B}([0,1])\big)$. This special construction will serve us well in many instances down the road, so it is worth studying in some detail.

When the distribution function $F(\cdot)$ is continuous and strictly increasing – that is, its graph has no jumps and no flat stretches – this construction is very easy: we just take $X(\omega) := F^{-1}(\omega)$, where $F^{-1}(\cdot)$ is the inverse of the given function $F(\cdot)$, and observe

$$\mathbb{P}(X \leq x) \;=\; \lambda\big(\{\omega \in \Omega \,|\, F^{-1}(\omega) \leq x\}\big) \;=\; \lambda\big([0, F(x)]\big) \;=\; F(x)\,, \qquad x \in \mathbb{R}\,.$$

Whenever F has discontinuities or flat stretches, the above procedure does not work, but can be modified appropriately; this modification is known as the SKOROHOD *construction.*

Proposition 6.3. SKOROHOD Construction: *For any given probability distribution function $F : \mathbb{R} \to [0,1]$, there exists a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and a random variable $X : \Omega \to \mathbb{R}$, such that $F_X(\cdot) \equiv F(\cdot)$.*

In particular, on the probability space $(\Omega, \mathcal{F}, \mathbb{P}) = ([0,1], \mathcal{B}([0,1]), \lambda|_{[0,1]})$ the mappings

$$X^+(\omega) := \inf\{x \,|\, F(x) > \omega\}\,, \quad X^-(\omega) := \inf\{x \,|\, F(x) \geq \omega\}\,, \quad 0 \leq \omega \leq 1 \tag{6.12}$$

or equivalently

$$X^+(\omega) = \sup\{x \,|\, F(x) \leq \omega\}\,, \quad X^-(\omega) := \sup\{x \,|\, F(x) < \omega\}\,,$$

are random variables with probability distribution functions $F_{X^\pm}(\cdot) \equiv F(\cdot)$. Furthermore, these mappings satisfy $\mathbb{P}(X^+ \neq X^-) = 0$.

Proof: We look at the right- and left-continuous inverses of $F(\cdot)$, namely the mappings $X^\pm$ of (1.3) on the space $(\Omega, \mathcal{F}, \mathbb{P}) \equiv ([0,1], \mathcal{B}([0,1]), \lambda)$ (draw a picture!), and notice the implications [37]

$$\omega \leq F(x) \;\Leftrightarrow\; X^-(\omega) \leq x\,, \qquad \omega < F(x) \;\Rightarrow\; X^+(\omega) \leq x\,,$$

[37] The second of these implications is fairly clear, as is the first half $\omega \leq F(x) \;\Rightarrow\; X^-(\omega) \leq x$ of the first. For the reverse of this implication, note that $z > X^-(\omega) \;\Rightarrow\; F(z) \geq \omega$, thus

$$F\big(X^-(\omega)\big) = \lim_{z \downarrow\downarrow X^-(\omega)} F(z) \geq \omega$$

thanks to the right-continuity of the function F, and therefore

$$X^-(\omega) \leq x \;\Rightarrow\; \omega \leq F\big(X^-(\omega)\big) \leq F(x)\,.$$

valid for all $\omega \in [0,1]$, $x \in \mathbb{R}$. These give

$$\mathbb{P}(X^- \le x) = \lambda\big(\{\omega \,|\, \omega \le F(x)\}\big) = F(x) \le \mathbb{P}(X^+ \le x)\,, \quad \forall\, x \in \mathbb{R}\,.$$

Of course $X^+ \ge X^-$, so $\{X^+ \ne X^-\} = \bigcup_{q\in\mathbf{Q}}\{X^- \le q < X^+\}$ where $\mathbf{Q}$ is the set of rationals. But we have

$$0 \le \mathbb{P}(X^- \le q < X^+) \;=\; \mathbb{P}\left(\{X^- \le q\} \setminus \{X^+ \le q\}\right) \;=\; F(q) - \mathbb{P}(X^+ \le q) \le 0\,, \quad \forall\, q \in \mathbb{R}\,,$$

thus $\mathbb{P}(X^+ \ne X^-) = 0$, because $\mathbf{Q}$ is countable. We conclude that $\mathbb{P}(X^+ \le x) = F(x)$ holds for all $x \in \mathbb{R}$. □

6.3 Probability Measures on Infinite-Dimensional Spaces

Let us tackle now the question of constructing measures on infinite-dimensional spaces.

We start with an infinite set $\mathfrak{T}$ (countable or not, it does not matter). For each $t \in \mathfrak{T}$ we let $\Omega_t = \mathbb{R}$, $\mathcal{F}_t = \mathcal{B}(\mathbb{R})$, and consider the *canonical* space

$$\Omega := \prod_{t\in\mathfrak{T}} \Omega_t \;\equiv\; \mathbb{R}^{\mathfrak{T}}$$

consisting of all real-valued functions $\omega : \mathfrak{T} \to \mathbb{R}$ on $\mathfrak{T}$.

We also consider the collection $\mathcal{C}^*$ of *finite-dimensional cylinder sets,* i.e., sets of the form

$$C \;=\; \big\{\omega \in \Omega \,|\, (\omega(t_1), \cdots, \omega(t_n)) \in A\big\} \quad \text{with } A \in \mathcal{B}(\mathbb{R}^n)\,,\;\; n \in \mathbb{N} \tag{6.13}$$

as well as the $\sigma-$algebra

$$\mathcal{F} \equiv \mathcal{B}\big(\mathbb{R}^{\mathfrak{T}}\big) := \sigma(\mathcal{C}^*)\,.$$

We denote by $\mathcal{T}_n$ the set of "finite sequences" $\tau = (t_1, \cdots, t_n)$, $n \in \mathbb{N}$, that is, of *distinct* $n-$tuples of elements in $\mathfrak{T}$, and set $\mathcal{T} := \bigcup_{n\in\mathbb{N}} \mathcal{T}_n$.

Suppose now that, for each $n \in \mathbb{N}$ and $\tau \in \mathcal{T}_n$, we have prescribed a probability distribution function $F_\tau : \mathbb{R}^n \to [0,1]$ with corresponding LEBESGUE-STIELTJES measure $\mu_\tau \equiv \mu_{F_\tau}$ on $\mathcal{B}(\mathbb{R}^n)$. We say that $\{F_\tau\}_{\tau\in\mathcal{T}}$ (respectively, $\{\mu_\tau\}_{\tau\in\mathcal{T}}$) is a family of *finite-dimensional probability distribution functions* (resp., of *finite-dimensional distributions*). Here is the question of interest:

Given a family $\{F_\tau\}_{\tau\in\mathcal{T}}$ as above, can we construct a probability measure $\mathbb{P}$ on $(\Omega, \mathcal{F})$ so that

$$\mathbb{P}[\,\omega \in \Omega \,|\, (\omega(t_1), \cdots, \omega(t_n)) \in A\,] \;=\; \mu_\tau(A)\,, \quad \forall\, A \in \mathcal{B}(\mathbb{R}^n) \tag{6.14}$$

holds for every $\tau = (t_1, \cdots, t_n) \in \mathcal{T}$ and $n \in \mathbb{N}$? In other words, can we put together a probability measure $\mathbb{P}$ on $\Omega = \mathbb{R}^{\mathfrak{T}}$ when we are given all its finite-dimensional "marginal" distributions $\{F_\tau\}_{\tau\in\mathcal{T}}$?

If such a measure $\mathbb{P}$ exists, then it is fairly straightforward to see from (6.14) that the following two **Consistency Conditions** *(C.C.'s)* have to be satisfied, for every $n \in \mathbb{N}$:

(C.C.1) PERMUTATION INVARIANCE: If $\varsigma = (t_{i_1}, \cdots, t_{i_n})$ is a permutation of $\tau = (t_1, \cdots, t_n) \in \mathcal{T}_n$, then for any Borel subsets $A_1, \cdots, A_n$ of the real line we have:

$$\mu_\tau \left(\prod_{j=1}^{n} A_j \right) = \mu_\varsigma \left(\prod_{j=1}^{n} A_{i_j} \right) .$$

(C.C.2) PARSIMONY: If $\tau = (t_1, \cdots, t_n) \in \mathcal{T}_n$ and $\varsigma = (t_1, \cdots, t_n, t_{n+1})$, then for any $B \in \mathcal{B}(\mathbb{R}^n)$ we have

$$\mu_\tau(B) = \mu_\varsigma(B \times \mathbb{R}) .$$

The following result asserts that these conditions are not only necessary, but also *sufficient* for the existence of such a probability measure $\mathbb{P}$. The proof uses in a crucial manner the regularity of each of the finite-dimensional distributions μ_τ, $\tau = (t_1, \cdots, t_n) \in \mathcal{T}_n$ on $\mathcal{B}(\mathbb{R}^n)$ for $n \in \mathbb{N}$, as in Exercise 4.26.

Theorem 6.2. DANIELL-KOLLMOGOROV: *Let $\{F_\tau\}_{\tau \in \mathcal{T}}$ be a given family of finite-dimensional p.d.f.'s, and suppose that the family of the corresponding finite-dimensional distributions $\{\mu_\tau\}_{\tau \in \mathcal{T}}$ satisfies the Consistency Conditions (C.C.1), (C.C.2) above.*

Then there exists a probability measure $\mathbb{P}$ on the canonical space $(\Omega, \mathcal{F})$ constructed above, such that (6.16) holds.

Example 6.2. Let $\{F_n\}_{n \in \mathbb{N}}$ be a sequence of probability distribution functions on the real line (with corresponding LEBESGUE-STIELTJES measures $\mu_n \equiv \mu_{F_n}$, $n \in \mathbb{N}$). We let $\mathfrak{T} = \mathbb{N}$, and consider the product probability distribution function

$$F_\tau(x_1, \cdots, x_n) := F_{t_1}(x_1) \cdots F_{t_n}(x_n) , \quad \forall\, (x_1, \cdots, x_n) \in \mathbb{R}^n$$

for any $n-$tuple $\tau = (t_1, \cdots, t_n) \in \mathcal{T}$ with associated LEBESGUE-STIELTJES measure

$$\mu_\tau \equiv \mu_{F_\tau} = \bigotimes_{j=1}^{n} \mu_{t_j}$$

on $\mathcal{B}(\mathbb{R}^n)$. It is clear that the family $\{\mu_\tau\}_{\tau \in \mathcal{T}}$ satisfies the Consistency Conditions of Theorem 6.2. According to this result, there exists a probability measure $\mathbb{P}$ on the canonical space $(\Omega, \mathcal{F}) \equiv (\mathbb{R}^{\mathbb{N}}, \sigma(\mathcal{C}^*))$ such that

$$\begin{aligned} \mathbb{P}\left[\, \omega \in \Omega \,|\, \omega(t_1) \in A_1 \,, \cdots, \omega(t_n) \in A_n \,\right] &= \mu_{t_1}(A_1) \cdots \mu_{t_n}(A_n) \qquad (6.15) \\ &= \prod_{j=1}^{n} \mathbb{P}\left[\, \omega \in \Omega \,|\, \omega(t_j) \in A_j \,\right] \end{aligned}$$

holds for any Borel subsets $A_1, \cdots, A_n$ of the real line, $n \in \mathbb{N}$, and $(t_1, \cdots, t_n) \in \mathcal{T}$. Under this probability measure, the coördinate mappings $X_n(\omega) := \omega_n$, $n \in \mathbb{N}$ are independent random variables with prescribed (one-dimensional marginal) distributions $F_n(x) = \mathbb{P}[\, X_n \leq x \,]$, $x \in \mathbb{R}$.

Proof of Theorem 6.2: For any cylinder set $C \in \mathcal{C}^*$ of the form (6.13), we set $\mathbb{P}(C) = \mu_\tau(A)$ where $\tau = (t_1, \cdots, t_n) \in \mathcal{T}_n$ and $A \in \mathcal{B}(\mathbb{R}^n)$. We leave it as an excercise, to check that the two consistency conditions (C.C.1), (C.C.2) guarantee $\mathbb{P}$ is well-defined and finitely-additive on $\mathcal{C}^*$ by this recipe, and $\mathbb{P}(\Omega) = 1$. If we can show that $\mathbb{P}$ is also *countably additive* on $\mathcal{C}^*$, then Theorem 6.1 will guarantee that $\mathbb{P}$ can be extended to a probability measure on $\mathcal{F} = \sigma(\mathcal{C}^*)$.

To this end, suppose that $\{B_k\}_{k\in\mathbb{N}}$ are disjoint sets in $\mathcal{C}^*$ with $B := \bigcup_{k\in\mathbb{N}} B_k \in \mathcal{C}^*$, set $C_m := B \setminus (\bigcup_{k=1}^m B_k)$ so that $\mathbb{P}(B) = \mathbb{P}(C_m) + \sum_{k=1}^m \mathbb{P}(B_k)$ holds for each $m \in \mathbb{N}$, and observe $\bigcap_{m\in\mathbb{N}} C_m = \emptyset$. Countable additivity will follow, as soon as we manage to show

$$\ell := \lim_{m\to\infty} \mathbb{P}(C_m) = 0\,. \tag{6.16}$$

The sequence $\{C_m\}_{m\in\mathbb{N}}$ is decreasing, so the limit in (6.16) exists. We shall assume that $\ell > 0$, and try to arrive at a contadiction.

Step 1: Monotonicity. We claim that, with this supposition, there exists a decreasing sequence $\{D_m\}_{m\in\mathbb{N}} \subset \mathcal{C}^*$ with the property $\bigcap_{m\in\mathbb{N}} D_m = \emptyset$ and $\lim_{m\to\infty} \mathbb{P}(D_m) = \ell > 0$, of the form

$$D_m = \{\omega \in \Omega \,|\, (\omega(t_1), \cdots, \omega(t_m)) \in A_m\} \quad \text{for some } A_m \in \mathcal{B}(\mathbb{R}^m)\,,$$

such that $\tau_m = (t_1, \cdots, t_m) \in \mathcal{T}_m$ is an extension of $(t_1, \cdots, t_{m-1}) \in \mathcal{T}_{m-1}$ for every $m \geq 2$.

To see this, observe that since $C_{k+1} \subseteq C_k$, each of the sets C_k is of the form: $C_k = \{\omega \in \Omega \,|\, (\omega(t_1), \cdots, \omega(t_{m_k})) \in A_{m_k}\}$ for some $A_{m_k} \in \mathcal{B}(\mathbb{R}^{m_k})$, with the properties that $A_{m_{k+1}} \subseteq A_{m_k} \times \mathbb{R}^{m_{k+1}-m_k}$ and $(t_1, \cdots, t_{m_{k+1}})$ is an extension of $(t_1, \cdots, t_{m_k})$. Define

$$D_1 = \{\omega \in \Omega \,|\, \omega(t_1) \in \mathbb{R}\}\ , \cdots, D_{m_1-1} = \left\{\omega \in \Omega \,|\, (\omega(t_1), \cdots, \omega(t_{m_1-1})) \in \mathbb{R}^{m_1-1}\right\}$$

and $D_{m_1} = C_1$; then we have

$$D_{m_1+1} = \{\omega \in \Omega \,|\, (\omega(t_1), \cdots, \omega(t_{m_1}), \omega(t_{m_1+1})) \in A_{m_1} \times \mathbb{R}\}\ , \qquad \cdots$$

$$D_{m_2-1} = \left\{\omega \in \Omega \,|\, (\omega(t_1), \cdots, \omega(t_{m_1}), \omega(t_{m_1+1}), \cdots, \omega(t_{m_2-1})) \in A_{m_1} \times \mathbb{R}^{m_2-m_1-1}\right\}$$

and $D_{m_2} = C_2$. Continuing this procedure, we see $\bigcap_{m\in\mathbb{N}} D_m = \bigcap_{m\in\mathbb{N}} C_m = \emptyset$.

Step 2: Regularity. From Exercise 4.26, there exists a closed set $F_m \subseteq A_m$ with the property $\mu_{\tau_m}(A_m \setminus F_m) < \varepsilon\, 2^{-m}$ for every $m \in \mathbb{N}$. Intersect this F_m with a sufficiently large closed ball to obtain a compact set K_m such that

$$E_m := \{\omega \in \Omega \,|\, (\omega(t_1), \cdots, \omega(t_m)) \in K_m\} \subseteq D_m\,, \quad \mathbb{P}(D_m \setminus E_m) = \mu_{\tau_m}(A_m \setminus K_m) < \frac{\varepsilon}{2^m}\,.$$

This sequence $\{E_m\}_{m\in\mathbb{N}}$ may not be decreasing, so we define $\widetilde{E}_m = \bigcap_{k=1}^m E_k$ and note that $\widetilde{E}_m = \{\omega \in \Omega \,|\, (\omega(t_1), \cdots, \omega(t_m)) \in \widetilde{K}_m\}$ with

$$\widetilde{K}_m = \left(K_1 \times \mathbb{R}^{m-1}\right) \bigcap \left(K_2 \times \mathbb{R}^{m-2}\right) \bigcap \cdots (K_{m-1} \times \mathbb{R}) \bigcap K_m$$

a compact set and $\mu_{\tau_m}(\widetilde{K}_m) = \mathbb{P}(\widetilde{E}_m) > 0$, because

$$\mathbb{P}(\widetilde{E}_m) \;=\; \mathbb{P}(D_m) - \mathbb{P}(D_m \setminus \widetilde{E}_m) \;=\; \mathbb{P}(D_m) - \mathbb{P}\left(\bigcup_{k=1}^{m}(D_m \setminus E_k)\right)$$

$$\geq\; \mathbb{P}(D_m) - \mathbb{P}\left(\bigcup_{k=1}^{m}(D_k \setminus E_k)\right) \;\geq\; \ell - \sum_{k=1}^{m}\frac{\ell}{2^k} \;>\; 0\,.$$

Step 3: Diagonalization. We have just shown that $\widetilde{K}_m$ is non-empty, so we may choose an element $(x_1^{(m)}, \cdots, x_m^{(m)}) \in \widetilde{K}_m$ for every $m \in \mathbb{N}$. The resulting sequence $\left\{x_1^{(m)}\right\}_{m\in\mathbb{N}}$ is contained in the compact set $\widetilde{K}_1$, so it must contain a subsequence $\left\{x_1^{(m_k)}\right\}_{k\in\mathbb{N}}$ that converges to some $x_1 \in \widetilde{K}_1$.

By the same token, $\left\{\left(x_1^{(m_k)}, x_2^{(m_k)}\right)\right\}_{k\in\mathbb{N}}$ is a sequence in the compact set $\widetilde{K}_2$, so it too contains a subsequence that converges to some $(x_1, x_2) \in \widetilde{K}_2$. Continuing this way we can put together a sequence of real numbers $(x_1, x_2, \cdots)$ such that $(x_1, x_2, \cdots, x_m) \in \widetilde{K}_m$ for each $m \in \mathbb{N}$. In other words,

$$S \;=\; \{\,\omega \in \Omega \,|\, \omega(t_i) = x_i\,,\; i \in \mathbb{N}\,\} \;\subset\; \widetilde{E}_m \;\subseteq\; D_m\,, \quad \forall\, m \in \mathbb{N}\,,$$

contradicting $\bigcap_{m\in\mathbb{N}} D_m = \emptyset$. This shows that (6.16) holds. □

7 Conditioning and Independence

Let us recall from section 2.9 the definition of conditional probability. For a fixed event F with $\mathbb{P}(F) > 0$ in a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, we define the conditional probability measure given F as

$$\mathbb{P}_F(E) \equiv \mathbb{P}(E \,|\, F) := \frac{\mathbb{P}(E \cap F)}{\mathbb{P}(F)}, \quad E \in \mathcal{F}. \tag{7.1}$$

This definition satisfies all the properties of a probability measure. We also have the following *Projectivity Property* of conditional probabilities: if $G \in \mathcal{F}$ also has positive probability, then

$$\mathbb{P}_F(E \,|\, G) = \mathbb{P}(E \,|\, F \cap G) = \mathbb{P}_{F\cap G}(E). \tag{7.2}$$

The proof is simple: from the definition (7.1) applied twice, we have

$$\mathbb{P}_F(E \,|\, G) = \frac{\mathbb{P}_F(E \cap G)}{\mathbb{P}_F(G)} = \frac{\mathbb{P}(E \cap G \cap F)}{\mathbb{P}(F)} \cdot \frac{\mathbb{P}(F)}{\mathbb{P}(F \cap G)} = \frac{\mathbb{P}(E \cap F \cap G)}{\mathbb{P}(F \cap G)} = \mathbb{P}(E \,|\, F \cap G).$$

Example 7.1. Royal Flush in Poker: A "royal flush" is a hand of five cards consisting of 10, J, Q, K, A in a single suit. You are a player, and receive 5 cards out of a total of 52 without been allowed to look at them; the probability that "you have received a royal flush" (event F) is

$$\mathbb{P}(F) = \frac{\#\text{ of royal flushes}}{\#\text{ of possible hands}} = 4 \cdot \frac{5!\,47!}{52!}.$$

Now suppose someone whispers into your ear, that she knows "you have received an ace of diamonds" (event A). *How has this new information changed your assessment of the chances to have received a royal flush?* Well, you compute now

$$\mathbb{P}(F \,|\, A) = \frac{\#\text{ of royal flushes with ace of diamonds}}{\#\text{ of possible hands with ace of diamonds}} = 1 \cdot \frac{4!\,47!}{51!} = \frac{13}{5} \cdot \mathbb{P}(F),$$

a probability significantly greater than $\mathbb{P}(F)$.

7.1 Partition Property and the BAYES Rule

The following *Partition Rule*

$$\mathbb{P}(E) = \mathbb{P}(E \,|\, F) \cdot \mathbb{P}(F) + \mathbb{P}(E \,|\, F^c) \cdot \mathbb{P}(F^c)$$

is also immediate from the definition (7.1); more generally, we have

$$\boxed{\mathbb{P}(E) = \sum_{j \in \mathcal{J}} \mathbb{P}(E \,|\, F_j) \cdot \mathbb{P}(F_j)} \tag{7.3}$$

for any (at most) countable partition $\Omega = \bigcup_{j\in\mathcal{J}} F_j$ of the sample space by disjoint events of positive probability.

From the partition property (7.3) we obtain for any event $E \in \mathcal{F}$ of positive probability the celebrated BAYES *Rule*

$$\boxed{\mathbb{P}(F_i \,|\, E) \;=\; \frac{\mathbb{P}(E \,|\, F_i) \cdot \mathbb{P}(F_i)}{\sum_{j \in \mathcal{J}} \mathbb{P}(E \,|\, F_j) \cdot \mathbb{P}(F_j)}\,, \qquad \forall\; i \in \mathcal{J}\,.} \tag{7.4}$$

Example 7.2. False Positives: *A disease strikes one in 100,000 people. A medical test is devised which gives a positive result with probability 0.95 when ill with the disease, and with probability 0.005 when not. How effective is the test in predicting the disease when positive?*

Denoting by F_1 (respectively, F_2) the event that one is (respectively, is not) ill with the disease, and by E the event that the test is positive, we obtain from (7.4) the computation

$$\mathbb{P}(F_1 \,|\, E) \;=\; \frac{\mathbb{P}(E \,|\, F_1) \cdot \mathbb{P}(F_1)}{\mathbb{P}(E \,|\, F_1) \cdot \mathbb{P}(F_1) + \mathbb{P}(E \,|\, F_2) \cdot \mathbb{P}(F_2)} \;=\; \frac{(0.95) \cdot (0.00001)}{(0.95) \cdot (0.00001) + (0.005) \cdot (0.99999)}$$

that is, approximately 0.002= 0.2% . The test is useless!

Example 7.3. GALTON's paradox: *We throw three coins: what is the probability they all turn out the same?*

Well, two will turn out the same anyway, so what's the probability the third will be the same as the other two? 50%, right?

Wrong. The sample space $\Omega = \{0,1\}^3$ contains eight elements, and the event $E = \{(000), (111)\}$ corresponding to "all coins turn out the same" has two elements. Taking $\mathcal{F} = \mathcal{P}(\Omega)$ with normalized counting measure (all eight possible outcomes equally likely) we obtain $\mathbb{P}(E) = 2/8 = 1/4$, which is the correct answer.

All right; what is wrong with the first "argument"? It is sloppy, too fast-and-loose. Let's try to make it rigorous: we shall use the partition rule (7.3) in this effort. The events

$$F_1 \;=\; \big\{(110),\, (101),\, (011),\, (111)\big\}\,, \qquad F_0 \;=\; \big\{(001),\, (010),\, (100),\, (000)\big\}$$

correspond to "at least two sides are heads" and to "at least two sides are tails", respectively. Each of them has probability equal to 1/2. On the other hand, the events $G_1 = \{(111)\}$ and $G_0 = \{(000)\}$ correspond to "all three sides are heads" and to "all three sides are tails", respectively, and from the definition of conditional probability we have

$$\mathbb{P}(E \,|\, F_1) \;=\; \frac{\mathbb{P}(E \cap F_1)}{\mathbb{P}(F_1)} \;=\; \frac{\mathbb{P}(G_1)}{\mathbb{P}(F_1)} \;=\; \frac{1}{4}\,, \qquad \mathbb{P}(E \,|\, F_0) \;=\; \frac{\mathbb{P}(E \cap F_0)}{\mathbb{P}(F_0)} \;=\; \frac{\mathbb{P}(G_0)}{\mathbb{P}(F_0)} \;=\; \frac{1}{4}$$

and thus (7.3) gives

$$\mathbb{P}(E) \;=\; \mathbb{P}(E \,|\, F_1) \cdot \mathbb{P}(F_1) + \mathbb{P}(E \,|\, F_0) \cdot \mathbb{P}(F_0) \;=\; \frac{1}{4} \cdot \frac{1}{2} + \frac{1}{4} \cdot \frac{1}{2} \;=\; \frac{1}{4}\,,$$

the same (correct) answer we got in the first place with much less effort.

Example 7.4. The MONTY HALL Paradox: This is a puzzle based on the television game show *"Let's Make a Deal"*. Its name comes from the show's host, Monty Hall; it is called a paradox because the result appears counterintuitive, yet is demonstrably true.

A well-known statement of the problem was published in *Parade magazine:* "Suppose you're on a game show, and you're given the choice of three doors: Behind one door is a car; behind the others, goats. You pick a door, say No. 1, and the host, who knows what's behind the doors, opens another door, say No. 3, which has a goat. He then says to you, *Do you want to pick door No. 2?* Is it to your advantage to switch your choice?"

As the player cannot be certain which of the two remaining unopened doors is the winning door, most people assume that each of these doors has an equal probability and conclude that switching does not matter. In fact, in the usual interpretation of the problem – where it is assumed that the player makes his first choice at random; that the host always shows a goat; and that, if he has a choice of doors to show, the host chooses at random – the player *should* switch; doing so doubles his probability of winning the car, from 1/3 to 2/3.

Why? Formula (7.3) again holds the key. Let us denote by A the event "win car", and by B the event "the right door was selected on the first try". Clearly, $\mathbb{P}(B) = 1/3$.
(i) If the player does not switch, his probability of winning is $\mathbb{P}(A) = \mathbb{P}(B) = 1/3$.
(ii) If he switches, then $\mathbb{P}(A|B) = 0$ and $\mathbb{P}(A|B^c) = 1$; since $\mathbb{P}(B^c) = 2/3$, his probability of winning is now

$$\mathbb{P}(A) \;=\; \mathbb{P}(A|B) \cdot \mathbb{P}(B) + \mathbb{P}(A|B^c) \cdot \mathbb{P}(B^c) \;=\; 2/3\,.$$

Example 7.5. College Admissions Paradox: A College receives a total of 3,150 applications, namely 2,083 from male (M) applicants and 1,067 from female (F) applicants. It ends up accepting (A) 1,341 applicants, namely 996 males and 345 females. Clearly

$$\mathbb{P}_M(A) \;=\; \mathbb{P}(A\,|\,M) \;=\; 47.81\%\;, \qquad \mathbb{P}_F(A) \;=\; \mathbb{P}(A\,|\,F) \;=\; 32.33\%$$

are the acceptance rates for male and female applicants, respectively, leading to immediate and vociferous accusations of bias. The Dean resigns in disgrace, and an Interim Dean is appointed who, as customary in such grave situations, appoints an *ad hoc* committee of distinguished faculty members, to investigate the matter and then make recommendations to the Trustees.

It occurs to someone on this committee to ask for acceptance/rejection data by discipline; and there a very puzzling picture emerges. The data are broken by school, or discipline $D_1,\, D_2,\, D_3,\, D_4$, as follows:

- D_1 received 825 applications from male candidates, out of which it accepted 511, or 62%; it received 108 applications from female candidates, out of which it accepted 85, or 79%.
- D_2 received 560 applications from male candidates, out of which it accepted 353, or 63%; it received 25 applications from female candidates, out of which it accepted 17, or 68%.
- D_3 received 325 applications from male candidates, out of which it accepted 110, or 34%; it received 593 applications from female candidates, out of which it accepted 219, or 37%.
- D_4 received 373 applications from male candidates, out of which it accepted 22, or 6%; it received 341 applications from female candidates, out of which it accepted 24, or 7%.

The committee cannot believe its eyes: how come the acceptance rates for females are higher in each and every one of the disciplines, yet the aggregate acceptance rate for male applicants is so

much higher than that for females? This cannot be; it is now the turn of the committee to resign, and the institution is brought to a standstill.

What is the resolution of this paradox? It comes in the form of (7.2) and (7.3), which suggest

$$\mathbb{P}_M(A) \,=\, \sum_{i=1}^{4} \mathbb{P}(A \,|\, D_i \cap M) \cdot \mathbb{P}_M(D_i)\,, \qquad \mathbb{P}_F(A) \,=\, \sum_{i=1}^{4} \mathbb{P}(A \,|\, D_i \cap F) \cdot \mathbb{P}_F(D_i)\,.$$

As we have observed, $\mathbb{P}(A \,|\, D_i \cap M) < \mathbb{P}(A \,|\, D_i \cap F)$ for every $i = 1, \cdots, 4$. But the probabilities $\mathbb{P}_M(D_i)$ and $\mathbb{P}_F(D_i)$ are not the same (compute them for each i), and their differences account for the inequality $\mathbb{P}_M(A) > \mathbb{P}_F(A)$, which goes in the reverse direction!

Exercise 7.1. Conditional Independence: Two events $A,\, B$ are said to be *conditionally independent,* given a third event C, if

$$\mathbb{P}\big(A \cap B \,|\, C\big) \,=\, \mathbb{P}\big(A \,|\, C\big) \cdot \mathbb{P}\big(B \,|\, C\big)\,.$$

Argue that this property is equivalent to

$$\mathbb{P}\big(A \,|\, C \cap B\big) \,=\, \mathbb{P}\big(A \,|\, C\big)\,.$$

Exercise 7.2. The LAPLACE "Law of Succession": An urn contains coins labelled $i = 1, \cdots, k$; we pick a coin "at random", meaning that each coin has probability $1/k$ to be selected. Once a selection as been made, the selected coin is flipped repeatedly, and the outcomes of different tosses are independent. We assume that coin i comes up heads with probability i/k on any given toss.

Suppose that the first n tosses of the selected coin have all come up heads; what is the probability that the next (the $(n+1)^{st}$) toss also comes up heads?

Exercise 7.3. Susan and Terry each toss a coin repeatedly and independently of each other, and count the number of failures (tails) they had to endure before seeing the first success (heads); say, S for Susan and T for Terry. (The two coins are assumed to be of similar manufacture.) After they are done, they announce the value of the total tally $S + T$.

What can we infer from this information, about the number of failures Susan had to endure? in other words, what are the conditional probabilities

$$\mathbb{P}\big(S = k \,|\, S + T = n\big)\,, \qquad k = 0, 1, \cdots, n$$

for a given $n \in \mathbb{N}$? State very precisely your assumptions.

7.2 Conditional Expectations

For a given event $F \in \mathcal{F}$ with $\mathbb{P}(F) > 0$ and any random variable $X : \Omega \to [0, \infty)$, we consider the expectation

$$\mathbb{E}(X \,|\, F) \,:=\, \mathbb{E}^{\mathbb{P}_F}(X) \,\equiv\, \int_\Omega X(\omega)\, d\mathbb{P}_F(\omega) \tag{7.5}$$

of the random variable with respect to the conditional probability measure of (7.1). This is called the *Conditional Expectation* of X given the event F.

It is clear that the quantity of (7.5) is

$$\mathbb{E}(X \mid F) \;=\; \frac{1}{\mathbb{P}(F)} \int_F X(\omega)\, \mathrm{d}\mathbb{P}(\omega) \;=\; \frac{1}{\mathbb{P}(F)} \int_\Omega X(\omega)\, \mathbf{1}_F(\omega)\, \mathrm{d}\mathbb{P}(\omega)\,.$$

It is also clear that for any (at most) countable partition $\Omega = \cup_{j\in\mathcal{J}} F_j$ of the sample space by disjoint events of positive probability, and for any integrable random variable $X : \Omega \to \mathbb{R}$, we have the so-called *Partition Rule for Expectations*

$$\boxed{\mathbb{E}(X) \;=\; \sum_{j\in\mathcal{J}} \mathbb{E}(X \mid F_j) \cdot \mathbb{P}(F_j)\,.} \tag{7.6}$$

For an indicator function of the form $X = \mathbf{1}_E$ with $E \in \mathcal{F}$, this property amounts to (7.3); so it holds for linear combinations of indicators, thus also for any $\mathcal{F}-$measurable $X : \Omega \to [0,\infty)$, and thence for any integrable X by arguments that are now familiar to us.

Exercise 7.4. A coin has probability $p \in (0,1)$ of coming up heads on any given toss. You keep tossing it, independently from time to time, until the first time you have seen it complete a run of k successive heads. How long will this take you, on the average? That is, if you denote by T the number of times you'll have to toss the coin until the first time you see this pattern, what is the expectation $\mathbb{E}(T) = \sum_{j\in\mathbb{N}} j\,\mathbb{P}(T=j)$ of this quantity? (*Hint:* Try the very easy case $k=1$ first, then generalize to an arbitrary natural number k.)

7.3 Independence and Product Measure

Let us recall the notion of Independence for Events and for Random Variables, from Definition 2.5. The following result is a corollary of the product measure and TONELLI-FUBINI Theorems 5.1-5.3.

Proposition 7.1. *On a given probability space $(\Omega, \mathcal{F}, \mathbb{P})$ consider two random variables $Y : \Omega \to \mathbb{R}$ and $Z : \Omega \to \mathbb{R}$, as well as the vector $(Y,Z) : \Omega \to \mathbb{R}^2$. The random variables Y, Z are independent if and only if*

$$\mu_{(Y,Z)} \;=\; \mu_Y \otimes \mu_Z\,, \tag{7.7}$$

and in this case

$$\begin{aligned}\mathbb{E}[\,\mathfrak{h}(Y,Z)\,] \;=\; \int_{\mathbb{R}^2} \mathfrak{h}(y,z)\,\mu_{(Y,Z)}(\mathrm{d}y,\mathrm{d}z) \;&=\; \int_{\mathbb{R}} \left(\int_{\mathbb{R}} \mathfrak{h}(y,z)\,\mu_Y(\mathrm{d}y) \right) \mu_Z(\mathrm{d}z) \\ &= \int_{\mathbb{R}} \left(\int_{\mathbb{R}} \mathfrak{h}(y,z)\,\mu_Z(\mathrm{d}z) \right) \mu_Y(\mathrm{d}y)\end{aligned}$$

holds for every BOREL *measurable function $\mathfrak{h} : \mathbb{R}^2 \to \mathbb{R}$ which is either nonnegative or satisfies $\mathbb{E}(|\mathfrak{h}(Y,Z)|) < \infty$.*

Indeed, if (7.7) holds, then for any sets $A \in \mathcal{B}(\mathbb{R})\,,\ B \in \mathcal{B}(\mathbb{R})$ we have $\mu_{(Y,Z)}(A \times B) = \mu_Y(A)\,\mu_Z(B)$, that is,

$$\mathbb{P}(Y \in A\,,\ Z \in B) \,=\, \mathbb{P}(Y \in A)\ \mathbb{P}(Z \in B)\,;$$

so Z and Y are then independent.

Conversely, if X and Y are independent, then (7.7) holds on the collection $\mathcal{R} = \{\, A \times B \,:\, A \in \mathcal{B}(\mathbb{R})\,,\ B \in \mathcal{B}(\mathbb{R})\,\}$ of (BOREL) measurable rectangles; but this collection is a $\pi-$system, and the two probability measures $\mu_{(Y,Z)}\,,\ \mu_Y \otimes \mu_Z$ agree on it, so they agree on the $\sigma-$algebra $\sigma(\mathcal{R}) = \mathcal{B}(\mathbb{R}) \otimes \mathcal{B}(\mathbb{R}) = \mathcal{B}(\mathbb{R}^2)$ by Theorem 4.8 and its Corollary.

It is fairly clear how to extend this corollary from the bivariate to the case of a vector $\mathfrak{Y} = (Y_1, \cdots, Y_d)$ with an arbitrary number of random variables.

Theorem 7.1. *Consider a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and on it independent random variables $Y_1, \cdots, Y_n$.*
(i) If these variables are integrable, then $\prod_{j=1}^n Y_j \in \mathbb{L}^1$ and we have

$$\boxed{\mathbb{E}\Big(\prod_{j=1}^n Y_j\Big) \,=\, \prod_{j=1}^n \mathbb{E}(Y_j)\,.} \tag{7.8}$$

(ii) If these variables are square-integrable and pairwise independent, then they are also pairwise uncorrelated and we have:

$$\boxed{\mathrm{Var}\Big(\sum_{j=1}^n Y_j\Big) \,=\, \sum_{j=1}^n \mathrm{Var}(Y_j)\,.} \tag{7.9}$$

Proof : (i) We shall deal only with $n = 2$; the general case requires only more complicated notation. With $f(y_1, y_2) = |y_1\, y_2|$ we have, from TONELLI's Theorem 5.2, Proposition 7.1, and the Composition and Change of Variable Theorem 4.5:

$$\mathbb{E}\Big(|Y_1 Y_2|\Big) \,=\, \int_{\mathbb{R}^2} f\ \mathrm{d}\big(\mu_1 \otimes \mu_2\big) \,=\, \int_{\mathbb{R}} |y_1|\,\mathrm{d}\mu_{Y_1}(y_1) \cdot \int_{\mathbb{R}} |y_2|\,\mathrm{d}\mu_{Y_2}(y_2) \,=\, \mathbb{E}(|Y_1|) \cdot \mathbb{E}(|Y_2|) < \infty\,.$$

We conclude that $Y_1\, Y_2 \in \mathbb{L}^1$; now we apply FUBINI's Theorem 5.3 (same argument, with absolute values removed).

For part (ii), observe that the random variables $\Xi_j := Y_j - \mathbb{E}(Y_j),\ j = 1, \cdots, n$ have zero expectation and are pairwise independent, thus $\mathbb{E}(\Xi_j \Xi_k) = \mathbb{E}(\Xi_j)\,\mathbb{E}(\Xi_k) = 0$ for $j \neq k$; therefore, the variables $Y_1, \ldots, Y_n$ are pairwise-uncorrelated, and

$$\mathrm{Var}\Big(\sum_{j=1}^n Y_j\Big) \,=\, \mathbb{E}\Big(\sum_{j=1}^n \Xi_j\Big)^2 \,=\, \sum_{j=1}^n \mathbb{E}(\Xi_j^2) + 2\sum_{j=1}^n \sum_{k=j+1}^n \mathbb{E}(\Xi_j \Xi_k) \,=\, \sum_{j=1}^n \mathrm{Var}(\Xi_j)\,. \qquad \square$$

As a consequence of this result, we have a very easy way to obtain the variance of the Binomial distribution in Exercise 4.4: we have $S_n = \sum_{i=1}^n X_i$ where $X_1, \cdots, X_n$ are independent

BERNOULLI variables, that is $\mathbb{P}(X_i = 1) = 1 - \mathbb{P}(X_i = 0) = p \in (0,1)$, so $\mathrm{Var}(X_i) = p$ and thus $\mathrm{Var}(S_n) = np$.

We are now in a position to state and prove the law of large numbers in its weak form. A version of this statement, in the case of simple coin-tossing, was already formulated by G. CARDANO, circa 1525; but the first proof appears in the *Ars Conjectandi* of BERNOULLI (1713). The generalization to independent, square integrable random variables stated here, is due to ČEBYŠEV (1867).

Theorem 7.2. The BERNOULLI (1713) / ČEBYŠEV (1867) Weak Law of Large Numbers: *On a probability space* $(\Omega, \mathcal{F}, \mathbb{P})$, *let* $X_1, X_2, \cdots$ *be independent, square-integrable random variables with common expectation* $m = \mathbb{E}(X_k)$ *and variances* $\sigma_k^2 = \mathrm{Var}(X_k)$ *that satisfy the condition* $\sum_{k=1}^n \sigma_k^2 = o(n^2)$ *as* $n \to \infty$. *Then with* $S_n = \sum_{k=1}^n X_k$ *we have*

$$\lim_{n\to\infty} \mathbb{P}\left(\left| \frac{S_n}{n} - m \right| > \varepsilon \right) = 0\,, \qquad \forall\ \varepsilon > 0\,. \tag{7.10}$$

We express this property by saying that the sequence of "sample averages" (S_n/n), $n \in \mathbb{N}$ *converges in probability* to the constant $m \in \mathbb{R}$; cf. chapter 10. Note that the conditions of the theorem are satisfied, if all the random variables $X_1, X_2, \cdots$ have the same variance; in particular, if they have the same distribution. As for the proof of (7.10), it is a straightforward consequence of the ČEBYŠEV inequality (5.4) and of (7.9), which guarantee that for every $\varepsilon > 0$ we have

$$\mathbb{P}\big(|S_n - \mathbb{E}(S_n)| > \varepsilon n\big) \le \frac{\mathrm{Var}(S_n)}{(\varepsilon n)^2} = \frac{1}{(\varepsilon n)^2} \sum_{k=1}^n \sigma_k^2 \longrightarrow 0\,, \quad \text{as } n \to \infty\,.$$

7.4 Convolution

Suppose that Y, Z are independent random variables with distributions $\mu \equiv \mu_Y$ and $\nu \equiv \mu_Z$, respectively, and $F(\cdot) = \mu((-\infty, \cdot])$, $G(\cdot) = \nu((-\infty, \cdot])$.

(i) The probability distribution function $H(x) = \mathbb{P}(X \le x)$, $x \in \mathbb{R}$ of the sum $X := Y + Z$ is given by the *convolution*

$$H(x) \equiv (F \star G)(x) := \int_{\mathbb{R}} F(x - z)\, \nu(\mathrm{d}z) = \int_{\mathbb{R}} G(x - y)\, \mu(\mathrm{d}y)\,, \quad x \in \mathbb{R} \tag{7.11}$$

of the distribution functions $F(\cdot)$ and $G(\cdot)$. This follows directly from Proposition 7.1, by taking $\mathfrak{h}(y,z) = \mathbf{1}_{\{y+z \le x\}}$.

Similarly, and equivalently, we have

$$\boxed{\mu_X(A) \equiv \big(\mu \star \nu\big)(A) = \int_{\mathbb{R}} \mu(A - z)\, \nu(\mathrm{d}z) = \int_{\mathbb{R}} \nu(A - y)\, \mu(\mathrm{d}y)\,,} \qquad A \in \mathcal{B}(\mathbb{R})\,. \tag{7.12}$$

Whereas, for any $\mathfrak{h} : \mathbb{R}^2 \to \mathbb{R}$ which is either nonnegative or such that $\mathbb{E}\big(|\mathfrak{h}(Y,Z)|\big) < \infty$, we have

$$\mathbb{E}\,\big(\mathfrak{h}(Y,Z)\big) = \int_{\mathbb{R}^2} \mathfrak{h}(y,z)\, \big(\mu \otimes \nu\big)(\mathrm{d}y, \mathrm{d}z) = \int_{\mathbb{R}} \left(\int_{\mathbb{R}} \mathfrak{h}(y,z)\, \mu(\mathrm{d}y) \right) \nu(\mathrm{d}z)$$
$$= \int_{\mathbb{R}} \left(\int_{\mathbb{R}} \mathfrak{h}(y,z)\, \nu(\mathrm{d}z) \right) \mu(\mathrm{d}y)\,.$$

(ii) As a consequence of (7.11) we have

$$\mathbb{P}(X = x) \;=\; \sum_{z\in\mathbb{R}} \mu(\{x-z\})\,\nu(\{z\}) \;=\; \sum_{y\in\mathbb{R}} \nu(\{x-y\})\,\mu(\{y\})\,, \qquad x \in \mathbb{R}.$$

Thus, if either F or G is continuous, we get $\mathbb{P}(X = x) = 0$ for every $x \in \mathbb{R}$, that is, H is then continuous as well.

(iii) If $F(x) = \int_{-\infty}^{x} f(u)\,du$ and $G(x) = \int_{-\infty}^{x} g(u)\,du$ are both diffuse with densities $f,\, g \in \mathbb{L}^+ \cap \mathbb{L}^1$, then $H \equiv F * G$ is also absolutely continuous, with density

$$h(x) \equiv (f \star g)(x) := \int_{-\infty}^{\infty} f(x-y)\,g(z)\,\mathrm{d}z = \int_{-\infty}^{\infty} g(x-y)\,f(y)\,\mathrm{d}y\,, \;\; x \in \mathbb{R} \tag{7.13}$$

given by the convolution of the two densities $f(\cdot)$ and $g(\cdot)$.

(iv) If Y has POISSON (λ_1) distribution, and Z has POISSON (λ_2) distribution, then $Y + Z$ has POISSON $(\lambda_1 + \lambda_2)$ distribution.

(v) If $Y_1, \cdots, Y_n$ are independent and exponentially distributed with the same parameter $\lambda > 0$, then $Y_1 + \cdots + Y_n$ has Gamma $\Gamma(\lambda, n)$ distribution.

For any $\lambda > 0,\, r > 0$ the *Gamma* $\Gamma(\lambda, r)$ *distribution* has probability density function

$$\lambda\, e^{-\lambda x}\; \frac{(\lambda x)^{r-1}}{\Gamma(r)}\,, \qquad x > 0\,, \tag{7.14}$$

where $\Gamma(r) := \int_0^\infty y^{r-1}\, e^{-y}\, dy$ is the classical Gamma function. The exponential is a special case of this distribution, corresponding to $r = 1$. When $r = n$ is an integer we have $\Gamma(n) = (n-1)!$ and this is the distribution of the sum of n independent exponential random variables with parameter $\lambda > 0$.

(vi) If Y has Gamma $\Gamma(\lambda, r)$ distribution and Z has Gamma $\Gamma(\lambda, s)$ distribution, then $Y + Z$ has Gamma $\Gamma(\lambda, r + s)$ distribution.

(vii) Suppose that a and b are given real numbers. If the random variable Y has Gaussian $\mathcal{N}(m_1, \sigma_1^2)$ distribution, and the random variable Z has Gaussian $\mathcal{N}(m_2, \sigma_2^2)$ distribution, then $aY + bZ$ has Gaussian $\mathcal{N}(a\,m_1 + b\,m_2, a^2\sigma_1^2 + b^2\sigma_2^2)$ distribution.

• The claims (iv)-(vii) are checked by direct computation. For instance, in the context of the claim made in (iv) we have

$$\big(\mu \otimes \nu\big)(\{n\}) \;=\; \sum_{k=0}^{n} \mu(\{k\})\,\nu(\{n-k\}) \;=\; \sum_{k=0}^{n} e^{-\lambda_1 k}\,\frac{\lambda_1^k}{k!} \cdot e^{-\lambda_2(n-k)}\,\frac{\lambda_2^{n-k}}{(n-k)!}$$

$$= \;\frac{e^{-(\lambda_1+\lambda_2)n}}{n!} \sum_{k=0}^{n} \frac{n!}{k!\,(n-k)!}\,\lambda_1^k\,\lambda_2^{n-k} \;=\; \frac{e^{-(\lambda_1+\lambda_2)n}}{n!}\,(\lambda_1 + \lambda_2)^n$$

for each $n \in \mathbb{N}_0$, from (7.12) and the Binomial theorem.

On the other hand, in the context of (v) let Y_1 and Y_2 be independent and have the same exponential distribution with parameter $\lambda > 0$. Then, according to (7.13) the distribution of $Y_1 + Y_2$ is diffuse, with probability density function that vanishes on $(-\infty, 0]$, whereas it equals

$$\int_{-\infty}^{\infty} \lambda\, e^{-\lambda(x-y)}\, \mathbf{1}_{[0,\infty)}(x-y) \cdot \lambda\, e^{-\lambda y}\, \mathbf{1}_{[0,\infty)}(y)\, dy \;=\; \int_0^x \lambda^2\, e^{-\lambda x}\, \mathrm{d}y \;=\; \lambda^2\, e^{-\lambda x}\, x\,, \qquad \text{for } x > 0\,.$$

This is the Gamma$(\lambda, 2)$ density of (7.14); the general form of the claim (v) is now obtained by induction and iteration of this argument.

Exercise 7.5. Beta-Gamma Calculus: Suppose that Y has Gamma $\Gamma(\lambda, r)$ distribution as in (7.14), that Z has Gamma $\Gamma(\lambda, s)$ distribution, and that Y and Z are independent.

(i) Show that $X := Y + Z$ has Gamma $\Gamma(\lambda, r+s)$ distribution.

(ii) Show that $B := Y/X$ has the so-called Beta (r, s) distribution, with probability density function

$$\frac{\Gamma(r+s)}{\Gamma(r)\,\Gamma(s)}\; x^{\,r-1}\,(1-x)^{\,s-1}\;, \qquad 0 < x < 1\,.$$

(i) Show that X and B are independent.

Exercise 7.6. Zeta Calculus: For $s > 1$ fixed, let $X : \Omega \to \mathbb{N}$ be a random variable with the "zeta distribution

$$\mathbb{P}(X = n) \;=\; \frac{1}{\zeta(s) \cdot n^s}\,, \quad n \in \mathbb{N}\,, \qquad \text{where } \; \zeta(s) \;:=\; \sum_{n \in \mathbb{N}} \frac{1}{n^s}$$

is the RIEMANN zeta-function. Consider for every given integer $m \in \mathbb{N}$ the event $A_m := \{\omega \in \Omega \,:\, m \text{ is a factor of } X(\omega)\}$.

(i) Show that $\mathbb{P}(A_m) = 1/m^s$.

(ii) Conclude from (i) that the events in the collection $\{A_p\}_{p \in \mathfrak{P}}$ are independent, where $\mathfrak{P}$ is the set of prime numbers.

(iii) Use (i) and (ii) to derive the EULER *formula*

$$\frac{1}{\zeta(s)} \;= \prod_{p \in \mathfrak{P}} \left(1 - \frac{1}{p^s}\right)\,.$$

Exercise 7.7. Skew representation of 2-D Gaussians: Suppose Y, Z are zero-mean Gaussian random variables with variances σ_1^2 and σ_2^2, respectively, and set $\sigma^2 := \sigma_1^2 + \sigma_2^2$.

(i) Show that the probability density function of the radius $R := \sqrt{Y^2 + Z^2}$ is given by

$$f_R(r) \;=\; \frac{r}{\sigma^2}\, \exp\left\{-\frac{r^2}{2\,\sigma^2}\right\}\,, \qquad r \in (0, \infty)\,.$$

(ii) Show that the angle $\Theta := \tan^{-1}\big(Z/Y\big)$ has uniform distribution on $[0, 2\pi]$, that is, probability density function $f_\Theta(\theta) = 1/(2\pi)$, $0 \le \theta < 2\pi$.

(i) Show that R and Θ are independent.

Exercise 7.8. For any two monotone increasing functions $f : \mathbb{R} \to \mathbb{R}$ and $g : \mathbb{R} \to \mathbb{R}$, and any random variable X on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, we have

$$\mathbb{E}\big[\, f(X)g(X)\,\big] \geq \mathbb{E}\big[\, f(X)\,\big] \cdot \mathbb{E}\big[\, g(X)\,\big]$$

if all expectations are well defined and finite. In particular, we have $\operatorname{Cov}\big(\, f(X),\, g(X)\big) \geq 0$.

7.5 Instances of Independence

Here are a few examples of situations, where independence arises quite naturally and, sometimes, unexpectedly or counterintuitively.

Example 7.6. RADEMACHER Functions: Let us consider the unit interval $\Omega = [0, 1)$ endowed with its BOREL sets and with LEBESGUE measure on this $\sigma-$alebra. It is well known that every number $\omega \in [0, 1)$ has a binary expansion

$$\omega \;=\; \frac{\varepsilon_1}{2} + \frac{\varepsilon_2}{2^2} + \cdots + \frac{\varepsilon_n}{2^n} + \cdots$$

where each ε is either 0 or 1.

In order to ensure the uniqueness of this expansion, we postulate that only expansions with infinitely many digits "0" are to be used. For instance, let us agree to write $\frac{3}{4}$ as $\frac{1}{2} + \frac{1}{2^2} + \frac{0}{2^3} + \frac{0}{2^4} + \cdots$ rather than $\frac{1}{2} + \frac{0}{2^2} + \frac{1}{2^3} + \frac{1}{2^4} + \cdots$; in other words, we prefer terminating expansions to nonterminating ones.

With this convention the digits $\varepsilon \in \{0, 1\}$ become functions of ω, so let us write more appropriately

$$\boxed{\omega \;=\; \frac{\varepsilon_1(\omega)}{2} + \frac{\varepsilon_2(\omega)}{2^2} + \cdots + \frac{\varepsilon_n(\omega)}{2^n} + \cdots}$$

or equivalently

$$\boxed{1 - 2\omega \;=\; \sum_{k\in\mathbb{N}} \frac{r_k(\omega)}{2^k}}, \qquad \text{where} \quad r_k(\omega) \;:=\; 1 - 2\,\varepsilon_k(\omega)\,, \qquad k \in \mathbb{N}$$

are the so-called RADEMACHER *functions*.

For instance, $\varepsilon_1(\omega) = 0$ for $\omega \in [0, 1/2)$ and $\varepsilon_1(\omega) = 1$ for $\omega \in [1/2, 1)$; similarly, $\varepsilon_2(\omega) = 0$ for $\omega \in [0, 1/4)$ or $\omega \in [1/2, 3/4)$, and $\varepsilon_2(\omega) = 1$ for $\omega \in [1/4, 1/2)$ or $\omega \in [3/4, 1)$; and so on. (Plot the first four functions in this sequence!)

Now let us follow KAC (1959), and do some simple trigonometry:

$$\begin{aligned}\sin x \;&=\; 2\sin\Big(\frac{x}{2}\Big)\cos\Big(\frac{x}{2}\Big) \;=\; 2^2\sin\Big(\frac{x}{4}\Big)\cos\Big(\frac{x}{4}\Big)\cos\Big(\frac{x}{2}\Big) \\ &= \cdots \;=\; 2^n \sin\Big(\frac{x}{2^n}\Big)\prod_{k=1}^{n}\cos\Big(\frac{x}{2^k}\Big)\end{aligned}$$

and note that $\lim_{n\to\infty} 2^n \sin(x2^{-n}) = x$, to obtain the *generalized* VIETA *formula*

$$\boxed{\frac{\sin x}{x} \;=\; \prod_{k\in\mathbb{N}} \cos\Big(\frac{x}{2^k}\Big)} \;. \tag{7.15}$$

But please observe

$$\int_0^1 e^{\,ix(1-2\omega)}\, d\omega \;=\; \frac{\sin x}{x}\,, \quad \int_0^1 \exp\left\{ ix\,\frac{r_k(\omega)}{2^k}\right\} d\omega \;=\; \frac{1}{2}\, e^{\,i(x/2^k)} + \frac{1}{2}\, e^{\,-i(x/2^k)} \;=\; \cos\Big(\frac{x}{2^k}\Big)\,,$$

so that the generalized VIETA formula can be written as

$$\int_0^1 \exp\left\{ ix \sum_{k\in\mathbb{N}} \frac{r_k(\omega)}{2^k}\right\} d\omega \;=\; \frac{\sin x}{x} \;=\; \prod_{k\in\mathbb{N}} \cos\Big(\frac{x}{2^k}\Big) \;=\; \prod_{k\in\mathbb{N}} \int_0^1 \exp\left\{ ix\,\frac{r_k(\omega)}{2^k}\right\} d\omega\,.$$

In particular, we get the RADEMACHER *formula*

$$\boxed{\int_0^1 \prod_{k\in\mathbb{N}} \exp\left\{ ix\,\frac{r_k(\omega)}{2^k}\right\} d\omega \;=\; \prod_{k\in\mathbb{N}} \int_0^1 \exp\left\{ ix\,\frac{r_k(\omega)}{2^k}\right\} d\omega\;.}$$

Here an integral of products is expressed as a product of integrals, much like the situation of Theorem 7.1.

Is this just simply a coincidence, or is it perhaps symptomatic of some underlying "independence" structure?

To make some headway with this question, let us endow $\Omega = [0,1)$ with its $\sigma-$algebra of BOREL sets $\mathcal{F} = \mathcal{B}([0,1))$ which measures ε_n for all $n \in \mathbb{N}$, and with LEBESGUE measure $\mathbb{P} \equiv \lambda$. Fix an arbitrary sequence $\mathfrak{d} = \{d_j\}_{j\in\mathbb{N}}$ of 0's and 1's, and look at the sets

$$I_j \;:=\; \{\,\omega\in\Omega\,|\,\varepsilon_j(\omega) = d_j\,\}\,, \qquad K_n \;:=\; \bigcap_{j=1}^{n} I_j \;:=\; \{\,\omega\in\Omega\,|\,\varepsilon_1(\omega) = d_1, \cdots, \varepsilon_n(\omega) = d_n\,\}\,;$$

this latter is the set of $\omega\in\Omega$, in whose binary expansion the first n digits are $d_1, \cdots, d_n$ and the rest are arbitrary:

$$\omega \;=\; \frac{d_1}{2} + \frac{d_2}{2^2} + \cdots + \frac{d_n}{2^n} + \frac{\varepsilon_{n+1}(\omega)}{2^{n+1}} + \frac{\varepsilon_{n+2}(\omega)}{2^{n+2}} + \cdots\;.$$

Clearly, K_n is an interval of length 2^{-n}, of the form

$$K_n \;=\; \Big[\,\sum_{j=1}^{n} d_j\, 2^{\,-j}\,,\; \sum_{j=1}^{n} d_j\, 2^{\,-j} + 2^{\,-n}\Big)\,. \tag{7.16}$$

Such intervals are called "dyadic", their endpoints being dyadic rationals with the same denominator; the number n is the *rank* or *order* of the interval. For each integer n the 2^n dyadic intervals of order n partition the unit interval. As we pass from the partition of rank n to that of rank $n+1$,

each closed/open interval (7.16) splits into two closed/open subintervals of equal length; the left half is assigned $\varepsilon_{n+1}(\omega) = 0$, whereas the right half gets $\varepsilon_{n+1}(\omega) = 1$.

Thus, informally speaking, we assign the value 0 to [0, 1/2), and the value 1 to [1/2, 1). We assign the value 00 to [0, 1/4); the value 01 to [1/4, 1/2); the value 10 to [1/2, 3/4); and the value 11 to [3/4,1). We assign the value 000 to [0, 1/8); the value 001 to [1/8, 1/4); the value 010 to [1/4, 3/8); the value 011 to [3/8, 1/2); the value 100 to [1/2, 5/8); the value 101 to [5/8, 3/4); the value 110 to [3/4, 7/8); and the value 111 to [7/8,1). And so on, as we already saw actually just a moment ago.

These considerations give us

$$\mathbb{P}\big(\{\,\omega \in \Omega \,|\, \varepsilon_j(\omega) = 0\,\}\big) \;=\; \mathbb{P}\big(\{\,\omega \in \Omega \,|\, \varepsilon_j(\omega) = 1\,\}\big) \;=\; 1/2\,, \qquad \forall\; j \in \mathbb{N}\,.$$

Since each I_j is a union of disjoint closed/open intervals with total length $\mathbb{P}(I_j) = 1/2$,

$$\mathbb{P}\big(\{\,\omega \in \Omega \,|\, \varepsilon_1(\omega) = d_1, \cdots, \varepsilon_n(\omega) = d_n\,\}\big) \;=\; \mathbb{P}(K_n) \;=\; 2^{-n} \;=\; \prod_{j=1}^{n}(1/2) \;=\; \prod_{j=1}^{n} \mathbb{P}(I_j)$$

$$= \prod_{j=1}^{n} \mathbb{P}\big(\{\,\omega \in \Omega \,|\, \varepsilon_j(\omega) = d_j\,\}\big)\,.$$

This is true for every choice of sequence $\mathfrak{d} = \{d_j\}_{j\in\mathbb{N}}$ of 0's and 1's, and integer n, and can be extended in a straightforward manner to

$$\mathbb{P}\big(\{\,\omega \in \Omega \,|\, \varepsilon_{k_1}(\omega) = d_1, \cdots, \varepsilon_{k_n}(\omega) = d_n\,\}\big) \;=\; \prod_{j=1}^{n} \mathbb{P}\big(\{\,\omega \in \Omega \,|\, \varepsilon_{k_j}(\omega) = d_j\,\}\big)$$

for any integers $1 \le k_1 < k_2 < \cdots < k_n$.

Therefore, the measurable functions $\varepsilon_1,\, \varepsilon_2\,, \cdots$ are independent "coin-tosses", that is, $\mathbb{P}(\varepsilon_j = 0) = \mathbb{P}(\varepsilon_j = 1) = 1/2$. This was first noticed by Émile BOREL in 1909.

But then the RADEMACHER functions $r_1,\, r_2\,, \cdots$ are also independent; they have the *"symmetric* BERNOULLI*" distribution*

$$\mathbb{P}(r_j = 1) \;=\; \mathbb{P}(r_j = -1) \;=\; 1/2\,.$$

Thus, the functions $\big\{e^{ix2^{-k}r_k}\big\}_{k\in\mathbb{N}}$ are independent as well, and the RADEMACHER formula becomes a special case of Theorem 7.1. □

Example 7.7. Inversions in Random Permutations: Let $\Omega \equiv \Sigma_n$ be the symmetric group of all $n!$ distinct permutations $\omega = (\omega_1, \cdots, \omega_n)$ of the integers $(1, \cdots, n)$. Consider the $\sigma-$algebra $\mathcal{F}$ of all subsets of this Ω, and denote by $\mathbb{P}$ the probability measure that assigns equal weight $1/n!$ to each such permutation.

For every $j \in \{1, \cdots, n\}$ and $\omega \in \Omega$, let $X_{nj}(\omega)$ be the number of *inversions* caused by j in ω: to wit, $X_{nj}(\omega) = k$ means that j precedes exactly k $(0 \le k \le j-1)$ of the integers $1, \cdots, j-1$ in the permutation ω. With this notation,

$$S_n(\omega) \;:=\; \sum_{j=1}^{n} X_{nj}(\omega)$$

is the total number of inversions in the permutation ω. For instance, with $n = 5$ the permutation $\omega = (3, 2, 5, 1, 4)$ of $(1, 2, 3, 4, 5)$ has $X_{n1}(\omega) = 0$, $X_{n2}(\omega) = 1$, $X_{n3}(\omega) = 2$, $X_{n4}(\omega) = 0$ and $X_{n5}(\omega) = 2$, thus $S_n(\omega) = 5$. We have then the following, rather remarkable, fact:

The random variables X_{nj}, $j = 1, \cdots, n$ *are independent, and*

$$\mathbb{P}\big(X_{nj} = k\big) \;=\; \frac{1}{j} \qquad \textit{for} \quad k = 0, \cdots, j-1\,.$$

Proof: The following argument is from CHUNG (1974): Let us start by observing that the values $X_{n1}(\omega), \cdots, X_{nj}(\omega)$ are determined, as soon as the sites occupied by the integers $1, \cdots, j$ in the permutation ω are known ("allotted"); the sites occupied by the remaining integers do not matter. Given j arbitrary sites among n ordered slots, there are $j!\,(n-j)!$ permutations ω in which the integers $1, \cdots, n$ occupy these sites in some order. Among these permutations, there are $(j-1)!\,(n-j)!$ permutations in which the integer j occupies the $(j-k)^{th}$ site, with the order from left to right, for some given $k \in \{0, 1, \cdots, j-1\}$. With this site fixed, there are $(j-1)!$ ways in which the integers $1, \cdots, j-1$ may occupy the remaining "allotted" sites; and each such way corresponds to *exactly one* of the possible values that the vector $(X_{n1}, \cdots, X_{n,j-1})$ can take.

Let us fix some such value $(c_1, \cdots, c_{j-1})$, and consider all permutations ω in which

(a) the integers $1, \cdots, j$ occupy the "allotted" sites; and

(b) $X_{n1}(\omega) = c_1\,, \cdots, X_{n,j-1}(\omega) = c_{j-1}\,,\;\; X_{nj}(\omega) = k\,.$

There are $(n-j)!$ such ω's; thus, the number of ω's that satisfy condition (b) is given by

$$\frac{n!}{j!\,(n-j)!} \cdot (n-j)! \;=\; \frac{n!}{j!}\,.$$

Now sum over $k \in \{0, 1, \cdots, j-1\}$ to find the number of ω's in which $X_{n1}(\omega) = c_1\,, \cdots,$ $X_{n,j-1}(\omega) = c_{j-1}$, namely: $j\,(n!/j!) \;=\; n!/(j-1)!$. Thus the claim follows from

$$\frac{\mathbb{P}[\,\omega \in \Omega \,:\, X_{n1}(\omega) = c_1\,, \cdots, X_{n,j-1}(\omega) = c_{j-1}\,,\;\; X_{nj}(\omega) = k\,]}{\mathbb{P}[\,\omega \in \Omega \,:\, X_{n1}(\omega) = c_1\,, \cdots, X_{n,j-1}(\omega) = c_{j-1}\,]} \;=\; \frac{n!}{j!} \cdot \frac{(j-1)!}{n!}. \;=\; \frac{1}{j}\,.$$

□

Example 7.8. Ranks and Records: Suppose $X_1, X_2, \cdots$ are independent random variables with *common* distribution function $F(\cdot)$ which is *continuous*. Consider the event

$$A_k := \Big\{ X_k > \max_{1 \le j \le k-1} X_j \Big\}$$

that "a record is set on day $t = k$", the number $W_n := \sum_{k=1}^{n} \mathbf{1}_{A_k}$ of records set by day $t = n$", as well as the random variable

$$R_n := 1 + \sum_{j=1}^{n-1} \mathbf{1}_{\{X_n < X_j\}}\,,$$

the relative rank of the random variable X_n among $X_1, \cdots, X_n$. Clearly, $A_n = \{R_n = 1\}$.

We claim that *the events* $\{A_n\}_{n\in\mathbb{N}}$ *are independent, with* $\mathbb{P}(A_n) = 1/n\,,\ n \in \mathbb{N}$. Thus, by the second BOREL-CANTELLI Lemma, we are going to be seeing a lot of reconrds getting broken: $\mathbb{P}(A_n,\text{ i.o.}) = 1$.

Even more to the point: *The random variables* $\{R_n\}_{n\in\mathbb{N}}$ *are also independent, with*

$$\mathbb{P}(R_n = \varrho) \,=\, \frac{1}{n}\,, \qquad \varrho = 1, \cdots, n\,.$$

Proof: We follow RESNICK (1999) and observe that, because $F(\cdot)$ is continuous, we have $\mathbb{P}(X_1 = X_2) = 0$, in fact $\mathbb{P}\big(\bigcup_{m\neq n}\{X_n = X_m\}\big) = 0$; consult section 7.4. (Observe that nothing here depends on the particular form of $F(\cdot)$, as long as this distribution function is continuous.) For fixed $n \in \mathbb{N}$, list the variables $X_1, \cdots, X_n$ in decreasing order

$$\max_{1\le i\le n} X_i \,=:\, Y_1^{(n)} \,>\, Y_2^{(n)} \,>\, \cdots > \, Y_n^{(n)} \,:=\, \min_{1\le i\le n} X_i$$

and define the random permutation $\pi^{(n)} = \big(\pi_1^{(n)} \cdots, \pi_n^{(n)}\big)$ of $(1, \cdots, n)$ as $\pi_i^{(n)} = r$ if $X_i = Y_r^{(n)}$; to wit, if the random variable X_i has relative rank r among $X_1, \cdots, X_n$, for $r = 1, \cdots, n$.

There are $n!$ permutations of $(1, \cdots, n)$, each of them corresponding to a particular ordering of the $X_1, \cdots, X_n$ (relative rankings $r_1, \cdots, r_n$). Every configuration $\{R_1 = \varrho_1, \cdots, R_n = \varrho_n\}$ determines uniquely an ordering of $X_1, \cdots, X_n$. For instance, with $n = 3$: the configuration $R_1(\omega) = 1,\ R_2(\omega) = 1,\ R_3(\omega) = 1,$ means $X_1(\omega) < X_2(\omega) < X_3(\omega)$; whereas $R_1(\omega) = 1,\ R_2(\omega) = 2,\ R_3(\omega) = 3,$ means $X_3(\omega) < X_2(\omega) < X_1(\omega)$.

Because of our assumptions, all such permutations are equally likely and we have

$$\mathbb{P}\big(\pi_1^{(n)} = \varrho_1, \cdots, \pi_n^{(n)} = \varrho_n\big) \,=\, \frac{1}{n!} \,=\, \mathbb{P}\big(R_1 = \varrho_1, \cdots, R_n = \varrho_n\big)\,.$$

In particular,

$$\mathbb{P}(R_n = \varrho_n) \,=\, \sum_{\varrho_1,\cdots,\varrho_{n-1}} \mathbb{P}\big(R_1 = \varrho_1, \cdots, R_n = \varrho_n\big) \,=\, \sum_{\varrho_1,\cdots,\varrho_{n-1}} \frac{1}{n!} \;\;;$$

each ϱ_j in this sum ranges over j values, so the number of terms in the sum is given by the product $1 \cdot 2 \cdots (n-1) = (n-1)!$. Therefore $\mathbb{P}(R_n = \varrho_n) = (n-1)!/\,n! = 1/n$, and

$$\mathbb{P}\big(R_1 = \varrho_1, \cdots, R_n = \varrho_n\big) \,=\, \frac{1}{n!} \,=\, \mathbb{P}\big(R_1 = \varrho_1\big) \cdots \mathbb{P}\big(R_n = \varrho_n\big)$$

for all possible values of $\varrho_j \in \{1, \cdots, j\}$ and $j = 1 \cdots, n$. The independence of the events $\{A_n\}_{n\in\mathbb{N}}$ follows now from that of the random variables $\{R_n\}_{n\in\mathbb{N}}$, since $A_n = \{R_n = 1\}$. $\square$

In particular, we get $\mathbb{P}(A_n) = \mathbb{P}(R_n = 1) = \mathbb{P}\big(\pi_n^{(n)} = 1\big) = 1/n$. This allows us to compute the mean and the variance of the number W_n of records set by day $t = n$, and their asymptotic behavior as $n \to \infty$:

$$\mathbb{E}(W_n) = \sum_{k=1}^{n}(1/k) \sim \log n \qquad \text{and} \qquad \text{Var}(W_n) = \sum_{k=1}^{n}(k-1)/k^2 \sim \log n\,.$$

Exercise 7.9. In the context of Example 7.6, consider the functions

$$Y(\omega) := \sum_{j\in\mathbb{N}} \frac{\varepsilon_{2j-1}(\omega)}{2^j}\,, \quad Z(\omega) := \sum_{j\in\mathbb{N}} \frac{\varepsilon_{2j}(\omega)}{2^j}\,, \quad W(\omega) := \sum_{j\in\mathbb{N}} \frac{2\,\varepsilon_j(\omega)}{3^j}\,.$$

Show that Y, Z are independent, with uniform distribution on the unit interval.

What can you say about the distribution function of W?

Example 7.9. The Inspection Paradox: Suppose that the bus arrives at (random) times S_1, $S_2\,\cdots$ at the stop close to your house, and that the "interarrival times" T_1, $T_2\,\cdots$ with $T_j := S_j - S_{j-1}$, $S_0 := 0$ are independent random variables with common exponential distribution given by $\mathbb{P}(T_j \geq u) = e^{-\lambda u}$ for $u \geq 0$ and $j \in \mathbb{N}$. Here λ is a positive constant, the reciprocal of the average interarrival time. This is the familiar setting of the POISSON process.

You arrive at the stop at a fixed time $\tau > 0$, say, exactly at noon; the time that has elapsed since the last bus came and went, and your waiting time until the next bus show up, are given as

$$X \,=\, \tau - S_{Q_\tau - 1}\,, \qquad Y \,=\, S_{Q_\tau} - \tau$$

respectively, where $Q_\tau := \inf\{\, n \in \mathbb{N} \mid S_n \geq \tau \,\}$. How long will you have to wait, on the average, before the bus arrives?

Heuristically, you might think that $\mathbb{E}(Y) = (1/2)\,\mathbb{E}(T_j) = 1/(2\lambda) = \mathbb{E}(X)$ should hold. This intuition is wrong; in fact, we have the following result:

The random variables X and Y are independent, with distributions

$$\mathbb{P}(X \geq x) \,=\, e^{-\lambda x}\,, \quad 0 \leq x < \tau \qquad \textit{and} \qquad \mathbb{P}(X = \tau) \,=\, e^{-\lambda \tau}\,, \tag{7.17}$$

and

$$\mathbb{P}(Y \geq y) \,=\, e^{-\lambda y}\,, \quad 0 \leq y < \infty\,. \tag{7.18}$$

In particular, the time until the next arrival has exponential distribution with parameter $\lambda > 0$, and from (5.22):

$$\mathbb{E}(Y) \,=\, \frac{1}{\lambda}\,, \qquad \mathbb{E}(X) \,=\, \int_0^\tau \mathbb{P}(X > x)\,\mathrm{d}x \,=\, \frac{1}{\lambda}\big(1 - e^{-\lambda\tau}\big) \,\approx\, \frac{1}{\lambda} \quad \text{for } \tau \text{ large}\,;$$

$$\mathbb{E}(X+Y) \,=\, \frac{1}{\lambda}\big(2 - e^{-\lambda\tau}\big) \,\approx\, \frac{2}{\lambda} \quad \text{for } \tau \text{ large}\,.$$

Proof: Indeed, for $0 \leq x \leq \tau$ and $y \geq 0$ we have

$$\mathbb{P}(X \geq x\,,\, Y \geq y) \,=\, \mathbb{P}\big(S_1 \geq \tau + y\big) + \sum_{j\in\mathbb{N}} \mathbb{P}\big(\tau - S_j \geq x\,,\, S_{j+1} - \tau \geq y\big)\,;$$

whereas for each fixed $j \in \mathbb{N}$ the random variables $T_{j+1} = S_{j+1} - S_j$ and S_j are independent, with distributions Exp(λ) and $\Gamma(j,\lambda)$, respectively, so that from Proposition 7.1 we have

$$\mathbb{P}\big(S_j \leq \tau - x\,,\, S_{j+1} \geq \tau + y\big) \,=\, \mathbb{P}\big(S_j \leq \tau - x\,,\, T_{j+1} \geq \tau + y - S_j\big)$$

$$= \int_0^{\tau-x} \left(\int_{\tau+y-s}^{\infty} \lambda\, e^{-\lambda u}\, du \right) \cdot \frac{\lambda^j}{\Gamma(j)}\, s^{j-1}\, e^{-\lambda s}\, \mathrm{d}s$$

$$= \int_0^{\tau-x} e^{-\lambda(\tau+y-s)} \cdot \frac{\lambda^j}{(j-1)!}\, s^{j-1}\, e^{-\lambda s}\, \mathrm{d}s$$

$$= \frac{\lambda^j}{(j-1)!}\, e^{-\lambda(\tau+y)} \int_0^{\tau-x} s^{j-1}\, \mathrm{d}s \;=\; \frac{(\lambda\,(\tau-x))^j}{j\,!}\, e^{-\lambda(\tau+y)}\,.$$

We deduce from this

$$\mathbb{P}(X \geq x\,,\, Y \geq y) \;=\; e^{-\lambda(\tau+y)} + \sum_{j\in\mathbb{N}} \frac{(\lambda\,(\tau-x))^j}{j\,!}\, e^{-\lambda(\tau+y)}$$

$$= \sum_{j\in\mathbb{N}_0} \frac{(\lambda\,(\tau-x))^j}{j\,!}\, e^{-\lambda(\tau+y)} \;=\; e^{-\lambda\,x}\, e^{-\lambda\,y}\,.$$

All claims follow now readily: The claim (7.18) upon setting $x = 0$; the claim (7.17) upon setting $y = 0$ and recalling $\mathbb{P}(X \geq \tau) = \mathbb{P}(X = \tau)$; whereas the independence is rather obvious. □

Discussion: The fact the the inter-arrival time $X + Y \;=\; S_{Q_\tau} - S_{Q_\tau - 1}$ that straddles the fixed "inspection time" τ, has expectation

$$\frac{1}{\lambda}\big(2 - e^{-\lambda\,\tau}\big) > \frac{1}{\lambda}\,,$$

constitutes the so-called *inspection paradox.*

I would have guessed that this difference $S_{Q_\tau} - S_{Q_\tau - 1}$ would have the same expectation as $S_k - S_{k-1} = T_{k-1}$, the typical inter-arrival time of the POISSON process. But you see, $S_{Q_\tau} - S_{Q_\tau - 1} = X + Y$ is *not* typical: by showing up at the fixed time τ, I introduce an "inspection bias" in favor of the longer waits in this POISSON process: I am far likelier to arrive during a long waiting interval between the arrivals of two successive buses, than during a short one (every time I see two M5 buses arrive in quick succession at the stop next to our house, I consider it an aberration or the result of poor coördination). The expected value of the bias thus introduced, is the difference $(1 - e^{-\lambda\,\tau})/\lambda$ between the quantities in the above display.

7.6 Constructing Sequences of Independent, Simple Random Variables

The procedure of Example 7.6 can be generalized greatly, to allow the construction of quite general sequence of independent random variables.

Theorem 7.3. *Let* $\{\mu_n\}_{n\in\mathbb{N}}$ *be a sequence of probability measures on the* BOREL *subsets of the real line, each with finite support. Then on a suitable probability space there exists a sequence* $X_1,\, X_2,\, \cdots$ *of independent, simple random variables with* $\mu_n = \mathbb{P} \circ X_n^{-1}\,,\;\; n \in \mathbb{N}\,.$

We shall take as probability space the unit interval $\Omega = [0,1)$ with $\mathcal{F} = \mathcal{B}([0,1))$ and $\mathbb{P}$ equal to LEBESGUE measure. We shall also denote by $\{\xi_1^{(n)}, \cdots, \xi_{K_n}^{(n)}\}$ the support of μ_n, so

$$\mu_n = \sum_{k=1}^{K_n} \mathfrak{p}_k^{(n)}\, \delta_{\xi_k^{(n)}} \qquad \text{for} \quad \mathfrak{p}_1^{(n)} > 0\,, \cdots, \mathfrak{p}_{K_n}^{(n)} > 0 \quad \text{and} \quad \sum_{k=1}^{K_n} \mathfrak{p}_k^{(n)} = 1\,.$$

Proof: (BILLINGSLEY (1986)) First, we partition $[0,1)$ into K_1 disjoint intervals $A_1^{(1)}, \cdots, A_{K_1}^{(1)}$ of respective lengths $\mathfrak{p}_1^{(n)}, \cdots, \mathfrak{p}_{K_n}^{(n)}$, and define the simple random variable

$$X_1(\omega) := \sum_{k=1}^{K_1} \xi_k^{(n)}\, \mathbf{1}_{A_k^{(1)}}(\omega)\,,$$

which has distribution μ_1, namely $\mathbb{P}(X_1 = \xi_k^{(1)}) = \mathbb{P}(A_k^{(1)}) = \mathfrak{p}_k^{(1)}\,,\ k = 1, \cdots, K_1$.

Next, we partition each of the above intervals $A_k^{(1)}$, $k = 1, \cdots, K_1$ into K_2 subintervals $A_{k1}^{(2)}, \cdots, A_{kK_2}^{(2)}$ of lengths $\mathfrak{p}_k^{(1)}\,\mathfrak{p}_1^{(2)}, \cdots, \mathfrak{p}_k^{(1)}\,\mathfrak{p}_{K_2}^{(2)}$, respectively; define the simple random variable

$$X_2(\omega) := \sum_{j=1}^{K_2} \xi_j^{(2)}\, \mathbf{1}_{\bigcup_{k=1}^{K_1} A_{kj}^{(2)}}(\omega)\,; \qquad \text{then} \quad \mathbb{P}(X_1 = \xi_k^{(1)}\,,\, X_2 = \xi_j^{(2)}) = \mathbb{P}(A_{kj}^{(2)}) = \mathfrak{p}_k^{(1)}\,\mathfrak{p}_j^{(2)}\,;$$

and adding up over k obtain $\mathbb{P}(X_2 = \xi_j^{(2)}) = \mathfrak{p}_j^{(2)}\,,\ j = 1, \cdots, K_2$. This leads to

$$\mathbb{P}\big(X_1 = \xi_k^{(1)}\,,\, X_2 = \xi_j^{(2)}\big) \;=\; \mathfrak{p}_k^{(1)}\,\mathfrak{p}_j^{(2)} \;=\; \mathbb{P}\big(X_1 = \xi_k^{(1)}\big) \cdot \mathbb{P}\big(X_2 = \xi_j^{(2)}\big)$$

for all $k = 1, \cdots, K_1$ and $j = 1, \cdots, K_2$, that is, the independence of X_1 and X_2.

Now we continue this procedure inductively: Suppose $[0,1)$ has been partitioned into $K_1 \cdot K_2 \cdots K_n$ disjoint intervals

$$A_{j_1 \cdots j_n}^{(n)}\,, \qquad\qquad j_1 = 1, \cdots, K_1\,, \quad \cdots \quad, \quad j_n = 1, \cdots, K_n \tag{7.19}$$

of respective lengths $\mathbb{P}\big(A_{j_1 \cdots j_n}^{(n)}\big) = \mathfrak{p}_{j_1}^{(1)} \cdots \mathfrak{p}_{j_n}^{(n)}$. We partition then the set $A_{j_1 \cdots j_n}^{(n)}$ into K_{n+1} subintevals $A_{j_1 \cdots j_n 1}^{(n+1)}, \cdots, A_{j_1 \cdots j_n K_{n+1}}^{(n+1)}$ of respective lengths

$$\mathbb{P}\big(A_{j_1 \cdots j_n}^{(n)}\big)\, \mathfrak{p}_1^{(n+1)}\,, \cdots, \mathbb{P}\big(A_{j_1 \cdots j_n}^{(n)}\big)\, \mathfrak{p}_{K_{n+1}}^{(n+1)}\,.$$

We repeat this procedure to create the next level $n+1$ of this cascade, and so on *ad infinitum.*

For an arbitrary level n of the resulting cascade, we define the random variable

$$X_n(\omega) \;:=\; \sum_{j=1}^{K_n} \xi_j^{(n)}\, \mathbf{1}_{\bigcup_{(k_1, \cdots, k_n)} A_{k_1 \cdots k_n j}^{(n)}}(\omega)\,,$$

where the union is over all indices $(k_1, \cdots, k_n)$ with $j_1 = 1, \cdots, K_1, \cdots, j_n = 1, \cdots, K_n$. Each partition (7.19) refines the one that precedes it in the cascade; we have

$$X_k(\omega) \,=\, \xi_{j_k}^{(k)} \qquad \text{for} \qquad \omega \in A_{j_1 \cdots j_n}^{(n)} \,,$$

so each element of (7.19) is contained in the set with the same labeling $j_1 \cdots j_n$ in the partition

$$\Lambda_{j_1 \cdots j_n} = \big\{\omega \in \Omega \,:\, X_1(\omega) = \xi_{j_1}^{(1)} \,, \cdots \,, X_1(\omega) = \xi_{j_1}^{(1)}\big\} \,, \quad j_1 = 1, \cdots, K_1 \,, \cdots \,, \; j_n = 1, \cdots, K_n \,.$$

As a consequence the two partitions coincide, and $\mathbb{P}\big(\Lambda_{j_1 \cdots j_n}\big) = \mathbb{P}\big(A_{j_1 \cdots j_n}^{(n)}\big) = \mathfrak{p}_{j_1}^{(1)} \cdots \mathfrak{p}_{j_n}^{(n)}$. Summing up over the indices $j_1, \cdots, j_{n-1}$ we see that X_n has distribution μ_n, and thus that the random variables $X_1, \cdots, X_n$ we constructed are independent. $\square$

7.7 BOREL, CANTELLI, and the First Strong Law of Large Numbers

The following two results, known as the first and second BOREL-CANTELLI Lemmata, respectively, are fundamental in the theory of Probability. It is a vast understatement, and only a result of historical accident, to call either one of these results a "Lemma".

Theorem 7.4. The BOREL (1909) and CANTELLI (1917) Lemmata: *For a sequence $\{E_n\}_{n \in \mathbb{N}} \subseteq \mathcal{F}$ of measurable sets in a measure space $(\Omega, \mathcal{F}, \mathbb{P})$, we have*

(i) $\mathbb{P}(E_n, \text{ i.o.}) = \mathbb{P}(\limsup_n E_n) = 0$, *if* $\sum_{n \in \mathbb{N}} \mathbb{P}(E_n) < \infty$.

(ii) *If* $\mathbb{P}(\Omega) = 1$, $\sum_{n \in \mathbb{N}} \mathbb{P}(E_n) = \infty$, *and the events* $\{E_n\}_{n \in \mathbb{N}}$ *are independent, then*

$$\mathbb{P}(E_n \,, \text{ i.o.}) \,=\, \mathbb{P}\Big(\limsup_{n \to \infty} \, E_n\Big) \,=\, 1$$

or equivalently $\sum_{n \in \mathbb{N}} \mathbf{1}_{E_n} = \infty$, *a.e.*

In other words, and in the context of a probability measure $\mathbb{P}(\Omega) = 1$: If the probabilities of the events decrease "very rapidly to zero", we cannot expect to "see too many of these events" being realized! Whereas, if the probabilities of *independent* events "do not decrease too rapidly to zero", we can expect to see these events realized infinitely often along a typical or "generic" realization $\omega \in \Omega$ (meaning, for every $\omega \in \Omega$ in some set $\Omega^* \in \mathcal{F}$ with $\mathbb{P}(\Omega^*) = 1$).

Proof of Theorem 7.4: (i) Recall from (4.6) that $\{E_n, \text{ i.o.}\} := \limsup_n E_n = \bigcap_{k \in \mathbb{N}} F_k$, where $F_k := \bigcup_{n \geq k} E_n$. Clearly, $\{F_k\}_{k \in \mathbb{N}}$ is a decreasing sequence and $\mathbb{P}(F_1) \leq \sum_{n \in \mathbb{N}} \mathbb{P}(E_n) < \infty$; likewise, $\mathbb{P}(F_k) \leq \sum_{n \in \mathbb{N},\, n \geq k} \mathbb{P}(E_n) < \infty$. Then the continuity-from-above property (4.14) implies $\mathbb{P}(\limsup_n E_n) = \mathbb{P}\left(\bigcap_{k \in \mathbb{N}} F_k\right) = \lim_{k \to \infty} \mathbb{P}(F_k) \leq \lim_{k \to \infty} \sum_{n \geq k} \mathbb{P}(E_n) = 0$.

(ii) On the other hand, for any $1 \leq k < m$, the independence of $\{E_n^c\}_{n=k}^m$ gives

$$0 \,\leq\, 1 - \mathbb{P}\left(\bigcup_{n=k}^m E_n\right) = \mathbb{P}\left(\bigcap_{n=k}^m E_n^c\right) = \prod_{n=k}^m \mathbb{P}(E_n^c) = \prod_{n=k}^m (1 - \mathbb{P}(E_n)) \leq \exp\left(-\sum_{n=k}^m \mathbb{P}(E_n)\right) \,,$$

where we have used the inequality $1-x \le e^{-x}$ for $0 \le x \le 1$. Since $\sum_{n\in\mathbb{N}} \mathbb{P}(E_n) = \infty$ we get, letting $m \to \infty$ and using the continuity-from-below property of the measure $\mathbb{P}$, that $1-\mathbb{P}(F_k) = 1-\mathbb{P}\big(\bigcup_{n\ge k} E_n\big) = 0$ holds for every $k \in \mathbb{N}$. But then the finiteness of the measure, along with the continuity-from-above property (4.14), give $\mathbb{P}(\limsup_n E_n) = \mathbb{P}\left(\bigcap_{k\in\mathbb{N}} F_k\right) = \lim_{k\to\infty}\mathbb{P}(F_k) = 1$, and (ii) is proved. □

It is hard to overstate how far one can go, by using just the BOREL-CANTELLI Lemmata intelligently and creatively. We illustrate the great usefulness of these results, by providing the first proof of the Strong law of Large Numbers, due to CANTELLI (1917).

Theorem 7.5. The CANTELLI (1917) Strong Law of Large Numbers: *Let $X_1, X_2, \cdots$ be a sequence of independent, zero-mean random variables with $\sup_{j\in\mathbb{N}} \mathbb{E}(X_j^4) \le K$ valid for some $K \in (0,\infty)$. Then we have*

$$\lim_{n\to\infty} \frac{1}{n}\sum_{j=1}^{n} X_j \,=\, 0\,, \qquad \textit{a.s.}$$

Proof of Theorem 7.5: Setting $S_n = \sum_{j=1}^n X_j$ we observe

$$\mathbb{E}\big(S_n^4\big) \,=\, \sum_{j=1}^{n} \mathbb{E}\big(X_j^4\big) + \sum_{\substack{j\ne i\\ j=1}}^{n}\sum_{i=1}^{n} \mathbb{E}\big(X_j^2 X_i^2\big) \,\le\, K\Big[\, n + n(n-1)\big)\,\Big] \,\le\, K\, n^2\,.$$

The reason is that terms of the form $X_i X_j X_k X_\ell$ with $i \ne j$ or $i \ne k$ or $i \ne \ell$ have zero expectation, because of the assumption of independence; on the other hand, we have also $\mathbb{E}(X_i^2) \le \sqrt{\mathbb{E}(X_i^4)} \le \sqrt{K}$ and $\mathbb{E}\big(X_j^2 X_i^2\big) = \mathbb{E}\big(X_j^2\big)\mathbb{E}\big(X_i^2\big) \le K$.

We would like to show that S_n/n tends to zero and n gets large; so we compare it to a sequence that decreases to zero as $n \to \infty$, such as $n^{-\alpha}$ for a suitable real constant $\alpha > 0$. Using ČEBYŠEV's inequality, we deduce from all this

$$\sum_{n\in\mathbb{N}} \mathbb{P}\left(\frac{|S_n|}{n} > \frac{1}{n^\alpha}\right) \le \sum_{n\in\mathbb{N}} \frac{\mathbb{E}\big(S_n^4\big)}{n^{4(1-\alpha)}} \le K \sum_{n\in\mathbb{N}} \frac{1}{n^{2(1-2\alpha)}} < \infty$$

for any given $\alpha \in (0, 1/4)$. From the first BOREL-CANTELLI Lemma, it follows now that the event

$$\left\{\omega \in \Omega \;\middle|\; \frac{|S_n(\omega)|}{n} > \frac{1}{n^\alpha} \text{ for infinitely many } n \in \mathbb{N}\right\}$$

has probability zero. This is to say, the event

$$\left\{\omega \in \Omega \;\middle|\; \frac{|S_n(\omega)|}{n} \le \frac{1}{n^\alpha} \text{ for all but finitely many } n \in \mathbb{N}\right\}$$

has probability equal to 1, and the result follows. □

We shall see vast generalizations of this result in Theorem 10.4. Additional illustrations of the BOREL-CANTELLI Lemmata can be found in the examples that follow.

A final point regarding Theorem 7.5: Its proof suggests that the speed of convergence to zero for the ratio $|S_n(\omega)|/n$, is at least $n^{-\alpha}$ for any $\alpha \in (0, 1/4)$. In fact, this ratio goes to zero considerably faster than that, in fact "very nearly like $1/\sqrt{n}$", in the sense $|S_n(\omega)|/n = O(n^{-\beta})$ for $0 < \beta < 1/2$. Two of the most important results in Probability Theory, the Central Limit Theorem and the Law of the Iterated Logarithm, substantiate this claim.

Example 7.10. Recurrence of Patterns: *What is the probability that in a sequence of independent coin tosses, say as in section 3.4, the pattern 100111 occurs infinitely often?*

Let us denote by A_k the event $\{X_k = 1,\, X_{k+1} = 0,\, X_{k+2} = 0,\, X_{k+3} = 1,\, X_{k+4} = 1,\, X_{k+5} = 1\}$ that such a run materializes starting at position k. The events $\{A_k\}_{k\in\mathbb{N}}$ have all the same probability, namely $p^4(1-p)^2$; they are not independent, but the events $\{A_{7j+1}\}_{j\in\mathbb{N}_0}$ are and we have clearly $\sum_{j\in\mathbb{N}_0} \mathbb{P}(A_{7j+1}) = \infty$. From Theorem 7.4 it follows that, with probability one, we shall see this pattern infinitely many times.

Needless to say, there is nothing special about this particular pattern; just replace it by any of your favorite texts, say ARISTOTLE's "Nicomachean Ethics", written in MORSE code.

Example 7.11. Runs in Coin Tossing: Suppose that $X_1, X_2, \cdots$ are independent Bernoulli random variables with $\mathbb{P}(X_j = 1) = \mathbb{P}(X_j = 0) = 1/2$ for all $j \in \mathbb{N}$. For any given $n \in \mathbb{N}$ we define $L_n = 0$ if $X_n = 1$; we define $L_n = k$ if $X_n = \cdots = X_{n+k-1} = 0$, $X_{n+k} = 1$; whereas we set $L_n = \infty$ if $X_{n+\ell} = 0$ for all $\ell \in \mathbb{N}_0$. To wit: L_n is the length of the "run of bad luck" (tails) that starts on day $t = n$.

We have clearly $\mathbb{P}(L_n \geq k) = \mathbb{P}(X_n = \cdots = X_{n+k-1} = 0) = 2^{-k}$ for $k \in \mathbb{N}_0$; in particular, $\mathbb{P}(L_n = \infty) = 0$. We claim

$$\mathbb{P}(L_n = 0\,,\ \text{i.o.}) = 1\,, \qquad \mathbb{P}(L_n = 1\,,\ \text{i.o.}) = 1\,. \tag{7.20}$$

Clearly, the events $\{L_n = 0\} = \{X_n = 1\}$, $n \in \mathbb{N}$ are independent and have probability 1/2 each, so the first claim follows from Theorem 7.4(ii). As for the events

$$A_n := \{L_n = 1\} = \{X_n = 0,\, X_{n+1} = 1\}\,, \quad n \in \mathbb{N}\,,$$

they all have the same probability $1/4$ but are of course not independent; however, the events $A_2, A_4, A_6, \cdots$ *are* independent, and for them Theorem 7.4(ii) gives $\mathbb{P}(L_{2n} = 1\,,\ \text{i.o.}) = 1$. The second claim in (7.20) follows from this.

- *Can we find a sequence* $\{a_n\}_{n\in\mathbb{N}} \subset (0,\infty)$ *such that* $\mathbb{P}\,(L_n \geq a_n\,,\ \text{i.o.}) = 0$ *?*

From Theorem 7.4(i), it is enough to have $\sum_{n\in\mathbb{N}} 2^{-a_n} < \infty$; this condition is satisfied by $a_n = (1+\varepsilon)\,\log_2 n$ for arbitrary $\varepsilon > 0$, so we obtain

$$\mathbb{P}\left(\frac{L_n}{\log_2 n} \geq 1+\varepsilon\,,\ \text{i.o.}\right) = 0\,, \qquad \text{thus} \qquad \limsup_{n\to\infty}\left(\frac{L_n}{\log_2 n}\right) \leq 1\,,\ \text{a.s.} \tag{7.21}$$

- *For every nondecreasing sequence* $\{b_n\}_{n\in\mathbb{N}} \subset (0,\infty)$ *with* $\sum_{n\in\mathbb{N}} (1/b_n)\, 2^{-b_n} = \infty$*, we have* $\mathbb{P}\,(L_n \geq b_n\,,\ \text{i.o.}) = 1$.

Indeed, consider such a sequence and define recursively $N_1 = 1$ and $N_{k+1} = N_k + b_{N_k}$ for $k \in \mathbb{N}$. This recipe is tailor-made to guarantee that the event

$$B_k := \{L_{N_k} \geq b_{N_k}\} = \{X_{N_k} = 0, \cdots X_{N_{k+1}-1} = 0\}$$

has probability $\mathbb{P}(B_k) = 2^{-b_{N_k}}$, and that $B_1, B_2, \cdots$ are *independent*. Thus, the claim will follow from Theorem 7.4(ii), as soon as we have shown that the series $\sum_{k\in\mathbb{N}} 2^{-b_{N_k}}$ diverges.

But this is now a straightforward consequence of the nondecrease of $\{b_n\}_{n\in\mathbb{N}}$ and of our assumptions, in conjunction with

$$\sum_{k\in\mathbb{N}} 2^{-b_{N_k}} = \sum_{k\in\mathbb{N}} \big(N_{k+1} - N_k\big)\, b_{N_k}^{-1}\, 2^{-b_{N_k}} \geq \sum_{k\in\mathbb{N}} \sum_{N_k \leq n < N_{k+1}} b_n^{-1}\, 2^{-b_n} = \sum_{n\in\mathbb{N}} b_n^{-1}\, 2^{-b_n}\,. \qquad \square$$

Now we just observe that the sequence $b_n = (1-\varepsilon)\,\log_2 n$ satisfies these conditions for any $0 \leq \varepsilon < 1$, so we obtain

$$\mathbb{P}\left(\frac{L_n}{\log_2 n} \geq 1\,,\ \text{i.o.}\right) = 1\,, \qquad \text{thus} \qquad \limsup_{n\to\infty}\left(\frac{L_n}{\log_2 n}\right) \geq 1\,,\ \text{a.s.}\,;$$

whereas, in conjunction with (7.21), this leads to the rather sharp asymptotics

$$\boxed{\limsup_{n\to\infty}\left(\frac{L_n}{\log_2 n}\right) = 1\,,\ \text{a.s.}}$$

Example 7.12. Growth of Exponential Variables: Suppose $X, X_1, X_2, \cdots$ are independent random variables with common exponential distribution $\mathbb{P}(X > u) = e^{-u}$, $u \geq 0$. Then the maximum-to-date sequence

$$M_n := \max_{1\leq k\leq n} X_k\,, \quad n \in \mathbb{N} \qquad \text{satisfies} \qquad M := \lim_{n\to\infty} M_n = \infty\,, \quad \text{a.e.}$$

Indeed, we have $\mathbb{P}(M \leq \xi) = \lim_{n\to\infty} \mathbb{P}(M_n \leq \xi) = \lim_{n\to\infty} \mathbb{P}(X_1 \leq \xi, \cdots, X_n \leq \xi) = \lim_{n\to\infty} \big(1 - e^{-\xi}\big)^n = 0$ for every $\xi \in (0,\infty)$, by independence.

Can we find how fast $(M_n)_{n\in\mathbb{N}}$ increases to infinity? In other words, can we identify a sequence of real numbers $\{a_n\}_{n\in\mathbb{N}} \subset (0,\infty)$ with $\lim_{n\to\infty} a_n = \infty$, such that

$$\lim_{n\to\infty}\big(M_n/a_n\big) = 1\,, \qquad \textit{holds a.e.?} \tag{7.22}$$

We claim that it is enough to identify this sequence, so that

$$\limsup_{n\to\infty}\big(X_n/a_n\big) = 1 \qquad \text{holds a.e.} \tag{7.23}$$

In other words, $\{a_n\}_{n\in\mathbb{N}}$ has to "increase fast enough", so that

$$\mathbb{P}\big(X_n > (1-\varepsilon)\,a_n\,,\ \text{i.o.}\big) = 1\,, \qquad \forall\ \varepsilon > 0$$

and therefore $\limsup_{n\to\infty}\big(X_n/a_n\big) \geq 1$ a.e.; but at the same time to "increase not too fast", so that we have also

$$\mathbb{P}\big(X_n > (1+\varepsilon)\, a_n \text{ , i.o.}\big) \,=\, 0\,, \qquad \forall\ \varepsilon > 0 \tag{7.24}$$

and thus $\limsup_{n\to\infty}\big(X_n/a_n\big) \leq 1$ a.e. The events $\{X_n > (1\mp\varepsilon)\, a_n\}$ under consideration are independent and have probabilities $e^{\,-(1\mp\varepsilon)\,a_n}$, respectively; thus, the BOREL-CANTELLI Lemmas guarantee both these statements if, for every $\varepsilon > 0$, we have

$$\sum_{n\in\mathbb{N}} e^{\,-(1-\varepsilon)\,a_n} \,=\, \infty \qquad \text{and} \qquad \sum_{n\in\mathbb{N}} e^{\,-(1+\varepsilon)\,a_n} \,<\, \infty\,.$$

Both these conditions, thus also (7.23), are clearly satisfied for the choice $a_n = \log n$.

• With this choice, let us justify the claim made above, to the effect that (7.23) implies (7.22) for the choice $a_n = \log n$. For every $\omega \in \Omega^*$ with $\mathbb{P}(\Omega^*) = 1$, it follows from (7.24) that there exists for each $\varepsilon > 0$ an integer $N_\varepsilon(\omega)$ such that $X_n(\omega) \leq (1+\varepsilon)\log n$, thus also

$$M_n(\omega) \,\leq\, M_{N_\varepsilon(\omega)}(\omega) + \max_{N_\varepsilon(\omega)<k\leq n} X_k(\omega) \,\leq\, M_{N_\varepsilon(\omega)}(\omega) + (1+\varepsilon)\log n\,,$$

holds for every $n > N_\varepsilon(\omega)$. This leads to the inequality $\limsup_{n\to\infty}\big(M_n(\omega)/\log n\big) \leq 1$.

In order to show that the inequality $\liminf_{n\to\infty}\big(M_n/\log n\big) \geq 1$ in the reverse direction also holds almost surely, it is enough to verify

$$\mathbb{P}\big(M_n < (1-\varepsilon)\,\log n\,,\ \text{i.o.}\big) \,=\, 0\,, \qquad \forall\ \varepsilon > 0\,.$$

But let us note that $\mathbb{P}(M_n < (1-\varepsilon)\,\log n) \,=\, (1-e^{\,-(1-\varepsilon)\,\log n})^n \,=\, (1-n^{\,\varepsilon-1})^n \,\leq\, e^{\,-n^\varepsilon}$ is the general term of a convergent series, so the claim follows once again from the first BOREL-CANTELLI lemma.

7.8 The KOLMOGOROV Zero-One Law

For any given sequence of random variables $X_1, X_2, \cdots$ let us denote by

- $\mathcal{F}_n := \sigma(X_1, \cdots, X_n)$ the smallest $\sigma-$algebra that measures the first $n \geq 1$ of them; by
- $\mathcal{T}^n := \sigma(X_{n+1}, X_{n+2}, \cdots)$ the smallest $\sigma-$algebra that measures all but the first $n \geq 0$ of them; and by
- $\mathcal{T} := \bigcap_{n\in\mathbb{N}_0} \mathcal{T}^n$ the *tail* or *remote* $\sigma-$algebra of this sequence.

Intuitively, the $\sigma-$algebra $\mathcal{T}$ contains all events whose occurrence is not affected by changing the values of finitely many terms in the sequence, and leaving all others the same. For instance, the event $\{X_n \in B_n, \text{ i.o. }\}$ belongs to $\mathcal{T}$, for any sequence $\{B_n\}_{n\in\mathbb{N}} \subset \mathcal{B}(\mathbb{R})$; so does the event $\{\lim_n (S_n/n) = 0\}$, where $S_n = \sum_{j=1}^n X_j$.

But events of the type $\{\, S_n = 0\,, \text{ i.o.}\}$ or $\{\, S_n > c_n\,, \text{ i.o.}\}$ do not belong to the $\sigma-$algebra $\mathcal{T}$: these belong to each σ-algebra $\sigma(S_n, S_{n+1}, \cdots)$, but not in each $\sigma(X_n, X_{n+1}, \cdots)$.

A celebrated result of KOLMOGOROV asserts that, for a sequence of *independent* random variables, the tail $\sigma-$algebra is trivial.

Theorem 7.6. KOLMOGOROV's Zero-One Law: *If the random variables $X_1, X_2, \cdots$ are independent, then $\mathcal{T} = \{\emptyset, \Omega\}$ mod. $\mathbb{P}$, that is:*

$$\mathbb{P}(A) \,=\, 0 \quad \textit{or} \quad 1\,, \qquad \textit{for every} \quad A \in \mathcal{T}\,.$$

Proof: Let us take any $A \in \mathcal{T}$ with $\mathbb{P}(A) > 0$ (clearly there is nothing to prove, if no such set exists); we shall try to show that $\mathbb{P}(A) = 1$. From Exercise 2.8 we know that $\mathcal{F}_n$ and $\mathcal{T}^n$ are independent for every $n \in \mathbb{N}$, so we have

$$\mathbb{P}(A \cap B) \,=\, \mathbb{P}(A) \cdot \mathbb{P}(B) \tag{7.25}$$

for every $B \in \mathcal{F}_n$ and $n \in \mathbb{N}$; this is because $A \in \mathcal{T} \subseteq \mathcal{T}^n$.

Thus (7.25) holds for every $A \in \mathcal{T}$ and $B \in \bigcup_{n\in\mathbb{N}} \mathcal{F}_n$. Note that

$$\mathcal{G} \,:=\, \bigcup_{n\in\mathbb{N}} \mathcal{F}_n$$

is a $\pi-$*system,* that is, closed under finite intersections: if $B_j \in \mathcal{F}_{n_j}$ for some $n_j \in \mathbb{N}$, $j = 1, 2$, then $B_1 \cap B_2 \in \mathcal{F}_n$ for $n := \max(n_1, n_2)$.

Now we define on $\mathcal{F}$ a new probability measure $\mathbb{P}_A(\cdot) := \mathbb{P}(A \cap \,\cdot\,)/\mathbb{P}(A)$; this is the "conditional probability measure given the event A as in (7.1). We observe that the two probability measures $\mathbb{P}_A$ and $\mathbb{P}$ agree on the $\pi-$system $\mathcal{G} = \bigcup_{n\in\mathbb{N}} \mathcal{F}_n$. Thus, by Corollary 4.1 these two probability measures agree on the σ-algebra

$$\sigma(\mathcal{G}) \,=\, \sigma\left(\bigcup_{n\in\mathbb{N}} \mathcal{F}_n\right) \,=\, \sigma(X_1, X_2, \cdots) = \mathcal{T}^0\,,$$

which contains $\mathcal{T}$. But this means that we can write (7.25) with $B = A$, namely $\mathbb{P}(A) = (\mathbb{P}(A))^2$, and leads to $\mathbb{P}(A) = 1$. $\square$

7.9 The HEWITT-SAVAGE Zero-One Law *

We go now one step farther, and consider the collection $\mathcal{E}$ of events in the σ-algebra $\sigma(X_1, X_2, \cdots)$ generated by a given sequence of random variables $X_1, X_2, \cdots$, which are permutation invariant: that is, not affected by permuting (re-arranging) a finite number $X_1, \cdots, X_n$ of these variables.

A bit more formally, let us call a one-to-one mapping $\mathbb{N} \ni j \mapsto \pi_j \in \mathbb{N}$ a *finite permutation*, if $\pi_j = j$ holds for all but finite many indices $j \in \mathbb{N}$. For the given sequence $\mathcal{X} = \{X_1, X_2, \cdots\}$ of random variables, and with a given set $A = \{\mathcal{X} \in B\},\ B \in \mathcal{B}(\mathbb{R}^{\mathbb{N}})$, we recall the construction of subsection 6.3 on the canonical space $\Omega = \mathbb{R}^{\mathbb{N}}$, and set $\pi(A) \equiv \{\pi(\mathcal{X}) \in B\}$.

We say that A is *permutation invariant,* if $\pi(A) = A$ holds for every finite permutation π of the natural numbers. The collection $\mathcal{E}$ of permutation-invariant sets, is thus

$$\mathcal{E} := \bigcap_{n \in \mathbb{N}} \mathcal{E}_n \,,$$

where $\mathcal{E}_n$ consists of the events that are invariant under permutations of the first n coördinates. Each $\mathcal{E}_n$ is a $\sigma-$algebra, thus so is $\mathcal{E}$; and we have clearly

$$\mathcal{T} \subseteq \mathcal{E} \,,$$

as "permuting" a finite number of coördinates is just one way of "changing" them.

For example, with $S_n = \sum_{j=1}^{n} X_j$, events of the type $\{S_n \in B, \text{ i.o.}\}$ for some BOREL set B, and $\{\limsup_{n\to\infty} S_n \geq 1\}$, belong to $\mathcal{E}$ (though the first of these need not belong to $\mathcal{T}$): re-arranging the order of a finite collection of random variables from the sequence $X_1, X_2, \cdots$, does not affect their sum.

The inclusion $\mathcal{T} \subseteq \mathcal{E}$ is of course *very* trivial, if the random variables $X_1, X_2, \cdots$ are independent; for then Theorem 7.6 asserts that $\mathcal{T} = \{\emptyset, \Omega\}$ mod. $\mathbb{P}$. The following result shows that if the $X_1, X_2, \cdots$ are not only independent but also identically distributed, we have actually $\mathcal{E} = \{\emptyset, \Omega\}$ mod. $\mathbb{P}$, as well: *both* of these σ-algebras are then trivial.

Theorem 7.7. HEWITT-SAVAGE Zero-One Law: *If the random variables $X_1, X_2, \cdots$ are independent and have the same distribution, then $\mathcal{E} = \{\emptyset, \Omega\}$ mod. $\mathbb{P}$, that is:*

$$\mathbb{P}(A) \,=\, 0 \quad \textit{or} \quad 1\,, \qquad \textit{for every} \quad A \in \mathcal{E}\,.$$

We shall not prove this result, but we send the reader to BILLINGSLEY (1986), p. 496 or SHIRYAEV (1989), p. 382 for concise arguments; see also DURRETT (2010).

8 Conditional Expectation Given a Sigma-Algebra

On a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ let us fix an event $F \in \mathcal{F}$ of positive measure $\mathbb{P}(F) > 0$, and recall from (7.1) the definition

$$\mathbb{P}_F(E) := \frac{\mathbb{P}(E \cap F)}{\mathbb{P}(F)}, \qquad E \in \mathcal{F} \tag{8.1}$$

of the *conditional probability* measure, given F. Integration with respect to this measure is denoted

$$\mathbb{E}_F(X) := \int_\Omega X \, d\mathbb{P}_F = \frac{1}{\mathbb{P}(F)} \cdot \int_F X \, d\mathbb{P} \tag{8.2}$$

for every random variable $X : \Omega \to \mathbb{R}$ with $\mathbb{E}(|X|) < \infty$. We can extend all these notions to the case $\mathbb{P}(F) = 0$, by setting then $\mathbb{P}_F(\cdot) \equiv 0$ on $\mathcal{F}$.

Now consider a (finite, or at most countable) partition $\Omega = \cup_{n=1}^N F_n$ of the space with $\{F_n\}_{n=1}^N \subseteq \mathcal{F}$ and $F_n \cap F_m = \emptyset$ for $m \neq n$. Countable additivity and (8.1), (8.2) imply

$$\mathbb{P}(E) = \sum_{n=1}^N \mathbb{P}(E \cap F_n) = \sum_{n=1}^N \mathbb{P}(F_n) \cdot \mathbb{P}_{F_n}(E), \qquad E \in \mathcal{F} \tag{8.3}$$

$$\mathbb{E}(X) = \sum_{n=1}^N \int_{F_n} X \, d\mathbb{P} = \sum_{n=1}^N \mathbb{P}(F_n) \cdot \mathbb{E}_{F_n}(X), \qquad X \in \mathbb{L}^1(\Omega, \mathcal{F}, \mathbb{P}). \tag{8.4}$$

Let us denote by $\mathcal{G}$ the smallest σ-algebra that contains $\{F_n\}_{n=1}^N$. For each $X \in \mathbb{L}^1(\Omega, \mathcal{F}, \mathbb{P})$ we define a simple function

$$\mathcal{X}(\omega) := \sum_{n=1}^N \mathbb{E}_{F_n}(X) \cdot \mathbf{1}_{F_n}(\omega), \qquad \omega \in \Omega \tag{8.5}$$

and note from (8.4) that its integral is $\mathbb{E}(\mathcal{X}) = \sum_{n=1}^N \mathbb{E}_{F_n}(X) \cdot \mathbb{P}(F_n) = \mathbb{E}(X)$. In fact we have something a little more general, namely

$$\mathbb{E}(X \cdot \mathbf{1}_\Lambda) = \mathbb{E}(\mathcal{X} \cdot \mathbf{1}_\Lambda), \qquad \forall\, \Lambda \in \mathcal{G}. \tag{8.6}$$

We shall denote this simple function $\mathcal{X}$ by $\mathbb{E}(X \,|\, \mathcal{G})$, and call it the *conditional expectation* of X, given the partition.

Exercise 8.1. Verify that the simple function $\mathcal{X} \equiv \mathbb{E}(X \,|\, \mathcal{G})$ of (8.5) is the unique (up to a.e. equivalence) $\mathcal{G}-$measurable function $H : \Omega \to \mathbb{R}$ with $\mathbb{E}(|H|) < \infty$ that satisfies

$$\mathbb{E}(X \cdot \mathbf{1}_\Lambda) = \mathbb{E}(H \cdot \mathbf{1}_\Lambda), \qquad \forall\, \Lambda \in \mathcal{G}. \tag{8.7}$$

Exercise 8.2. Let $\Omega = \{(\omega_1, \omega_2) \,|\, \omega_i = 1, \ldots, 6 \text{ for } i = 1, 2\}$ consist of the 36 possible outcomes when tossing a die twice, let $\mathcal{F} = 2^\Omega$ be the collection of all subsets of Ω, and define $X(\omega) = \omega_1$, $Y(\omega) = \omega_1 + \omega_2$.

Considering the partition $F_n = \{\, \omega \in \Omega \,|\, Y(\omega) = n \,\}$ for $n = 2, \ldots, 12$, compute $\mathbb{E}_{F_n}(X)$ for each n and verify the property $\mathbb{E}\big(\mathbb{E}(X \,|\, \mathcal{G})\big) = \mathbb{E}(X)$.

Actually, a somewhat closer look at (8.6) suggests a more general approach. Suppose that we have an arbitrary sub-σ-algebra $\mathcal{G}$ of events in $\mathcal{F}$ and consider a random variable $X : \Omega \to [0, \infty)$ with $\mathbb{E}(X) < \infty$. Then the set function

$$\mathbb{Q}_X(\Lambda) := \mathbb{E}(X \cdot \mathbf{1}_\Lambda) = \int_\Lambda X \, \mathrm{d}\mathbb{P}, \qquad \Lambda \in \mathcal{G}$$

on the right-hand side of (8.6) is non-negative, countably-additive and finite, since $\mathbb{Q}_X(\Omega) = \mathbb{E}(X) < \infty$. It is also absolutely continuous with respect to $\mathbb{P}$.

From the Radon-Nikodým Theorem 5.5 we conclude that there exists a unique (up to a.e. equivalence) $\mathcal{G}-$measurable function $H := (d\mathbb{Q}_X / d\mathbb{P})$ from Ω into $[0, \infty)$, such that $\mathbb{Q}_X(\Lambda) = \int_\Lambda H \, d\mathbb{P}$, or equivalently $\mathbb{E}(X \cdot \chi_\Lambda) = \mathbb{E}(H \cdot \chi_\Lambda)$, holds for all $\Lambda \in \mathcal{G}$; in particular, this gives $\mathbb{E}(H) = \mathbb{E}(X) < \infty$. We shall denote this function by $\mathbb{E}(X \,|\, \mathcal{G})$, as indicated below (8.6).

More generally, if we are given an arbitrary integrable random variable $X : \Omega \to \mathbb{R}$, we can repeat this procedure to its positive and negative parts X^+ and X^- and come up with an integrable, $\mathcal{G}-$measurable random variable

$$\mathcal{X} := \mathbb{E}(X^+ \,|\, \mathcal{G}) - \mathbb{E}(X^- \,|\, \mathcal{G})$$

that satisfies (8.6). Again, we shall denote this random variable $\mathcal{X}$ by $\mathbb{E}(X \,|\, \mathcal{G})$.

These remarks lead us to the following, formal approach.

Definition 8.1. Conditional Expectation given a sigma-Algebra: Let $\mathcal{G}$ be a sub-σ-algebra of events in a complete probability space $(\Omega, \mathcal{F}, \mathbb{P})$. For any given random variable $X : \Omega \to \mathbb{R}$ with $\mathbb{E}(|X|) < \infty$, there exists a unique (up to a.e. equivalence) $\mathcal{G}-$measurable and integrable random variable $\mathbb{E}(X \,|\, \mathcal{G}) : \Omega \to \mathbb{R}$, called the *conditional expectation of* X *given* $\mathcal{G}$, that satisfies

$$\mathbb{E}\big(X \cdot \mathbf{1}_\Lambda\big) = \mathbb{E}\Big(\mathbb{E}(X \,|\, \mathcal{G}) \cdot \mathbf{1}_\Lambda\Big), \qquad \forall\, \Lambda \in \mathcal{G}. \tag{8.8}$$

I cannot forget how surprised I was when, as a graduate student just like you, I was told for the first time to think of conditional expectation as a random variable! But I should not have been. For example, it should have occurred to me that life expectancy at birth *conditional on sex* is a random variable, that takes one value for each make and another value for each female. [38]

Definition 8.2. Conditional Probability given a sigma-Algebra: In the setting of Definition 8.1 there exists for any given $E \in \mathcal{F}$ a unique (up to a.e. equivalence) $\mathcal{G}-$measurable random variable $\mathbb{P}(E \,|\, \mathcal{G}) : \Omega \to [0, 1]$, called the *conditional probability of* E *given* $\mathcal{G}$, that satisfies

$$\mathbb{P}\big(E \cap \Lambda\big) = \mathbb{E}\Big(\mathbb{P}(E \,|\, \mathcal{G}) \cdot \mathbf{1}_\Lambda\Big), \qquad \forall\, \Lambda \in \mathcal{G}. \tag{8.9}$$

[38] This is simple and intuitive enough; but taken to its logical conclusion, it has very important consequences. For instance, with continuous random variables, we often have to condition on an event of probability zero; how exactly to do this, had bedeviled mathematicians for centuries. But considering arbitrary sigma algebras, rather than only partitions consisting of sets of positive measure, does that seamlessly; see Proposition 8.6 and Remarks 8.4, 8.5.

As it turns out, with the publication of the OTTO NIKODÝM paper in 1930, KOLMOGOROV understood immediately that he finally had the missing piece he needed, to complete the rigorous treatment and "mathematization" of the field, that became his magnum opus KOLMOGOROV (1933).

Remark 8.1. The requirement (8.8) is clearly satisfied, if we have

$$\mathbb{E}\big(X \cdot \Xi\big) \,=\, \mathbb{E}\big(\,\mathbb{E}(X\,|\,\mathcal{G}) \cdot \Xi\big)\,, \quad \forall\ \Xi \in \mathbb{L}^\infty(\Omega, \mathcal{G}, \mathbb{P})\,. \tag{8.10}$$

Building our way up, from indicators to simple functions to bounded measurable functions, we can show that (8.10) is actually implied by (8.8), thus equivalent to it.

Remark 8.2. When $\mathcal{G} \,=\, \sigma(\Xi_\alpha\,,\ \alpha \in A)$ is the σ-algebra generated by a family of random variables, we write suggestively $\mathbb{E}\,(X\,|\,\Xi_\alpha\,,\ \alpha \in A)$ for $\mathbb{E}(X\,|\,\mathcal{G})$. Similarly for conditional probabilities.

Proposition 8.1. Conditional Expectation as Projection: *If X is square-integrable, then so is its conditional expecation $\mathbb{E}(X\,|\,\mathcal{G})$ given any sub-σ-algebra $\mathcal{G} \subseteq \mathcal{F}$; this conditional expectation minimizes the $\mathbb{L}^2$ distance of X*

$$\mathbb{E}\big(X - \mathbb{E}(X\,|\,\mathcal{G})\big)^2 \,=\, \inf_{Y \in \mathbb{L}^2(\mathcal{G})} \mathbb{E}(X - Y)^2 \tag{8.11}$$

from elements in the space $\mathbb{L}^2(\mathcal{G})$ of square-integrable, $\mathcal{G}-$measurable random variables.

Proof: The space $\mathbb{L}^2(\mathcal{F}) \equiv \mathbb{L}^2(\Omega, \mathcal{F}, \mathbb{P})$ is a Hilbert space with inner product $\langle X, Y\rangle \,\equiv\, \mathbb{E}(XY)$, and $\mathbb{L}^2(\mathcal{G})$ is a closed subspace in $\mathbb{L}^2(\mathcal{F})$. From the orthogonal projection Theorem 16.1 there exists a unique element H in $\mathbb{L}^2(\mathcal{G})$ such that (8.11) holds, as well as

$$\langle X - H, \Xi\,\rangle \,=\, 0\,, \qquad \forall\ \Xi \in \mathbb{L}^2(\mathcal{G})\,,$$

or equivalently $\mathbb{E}(X\,\Xi\,) \,=\, \mathbb{E}(H\,\Xi\,)$. Comparing with (8.8) we see that H is indeed (a version of) the conditional expectation $\mathbb{E}(X\,|\,\mathcal{G})$. □

Remark 8.3. The interpretation of conditional expectation as projection helps motivate the following property, which can be phrased as saying that: *if you are projecting first on a large subspace and then down to a smaller one, you might as well project directly to the smaller subspace.*

Proposition 8.2. Tower Property of Conditional Expectation *If X is integrable and $\mathcal{G}_1 \subseteq \mathcal{G}_2$ are sub-σ-algebras of $\mathcal{F}$, then*

$$\mathbb{E}\big(\mathbb{E}\big(X\,|\,\mathcal{G}_2\big)\,|\,\mathcal{G}_1\big) \,=\, \mathbb{E}\big(X\,|\,\mathcal{G}_1\big)\,, \quad \text{a.s.} \tag{8.12}$$

Proof: An arbitrary $\Lambda \in \mathcal{G}_1$ also belongs to $\mathcal{G}_2$, so the defining property (8.8) – used first for $\mathcal{G}_2$, then for $\mathcal{G}_1$ – gives $\mathbb{E}\big(\,\mathbb{E}(X\,|\,\mathcal{G}_2) \cdot \mathbf{1}_\Lambda\big) \,=\, \mathbb{E}(X \cdot \mathbf{1}_\Lambda) \,=\, \mathbb{E}\big(\mathbb{E}(X\,|\,\mathcal{G}_1) \cdot \mathbf{1}_\Lambda\big)$. But $\mathbb{E}(X\,|\,\mathcal{G}_1)$ is $\mathcal{G}_1-$measurable, so (8.12) follows from yet another application of (8.8). □

Proposition 8.3. "Taking out what is known": *Let $\mathcal{G}$ be a sub-σ-algebra of $\mathcal{F}$.*
If X is $\mathcal{F}-$measurable, Y is $\mathcal{G}$-measurable, and $\mathbb{E}(|X|) \,+\, \mathbb{E}(|XY|) \,<\, \infty$, then we have

$$\mathbb{E}(XY\,|\,\mathcal{G}) \,=\, Y \cdot \mathbb{E}\,(X\,|\,\mathcal{G})\,, \quad \text{a.s.} \tag{8.13}$$

Proof: For every $\Lambda \in \mathcal{G}$ we need to show

$$\mathbb{E}\,(XY \cdot \mathbf{1}_\Lambda) \;=\; \mathbb{E}\,(\,\mathbb{E}\,(X \,|\, \mathcal{G}) \cdot Y\,\mathbf{1}_\Lambda\,)\,. \tag{8.14}$$

This is clear if $Y \in \mathbb{L}^\infty(\mathcal{G})$, because then $\Xi := Y \cdot \mathbf{1}_\Lambda$ is also in $\mathbb{L}^\infty(\mathcal{G})$ and (8.14) follows from (8.10). If Y is not necessarily bounded but both X and Y are nonnegative, we can write $Y = \lim_{n\to\infty} \uparrow\, Y_n$ as the increasing limit of a sequence $\{Y_n\}_{n\in\mathbb{N}} \subseteq \mathbb{L}^\infty(\mathcal{G})$ of simple nonnegative $\mathcal{G}-$measurable functions, then use the Monotone Convergence Theorem to justify

$$\mathbb{E}\big(X \cdot Y\,\mathbf{1}_\Lambda\big) \;=\; \lim_{n\to\infty} \mathbb{E}\big(X \cdot Y_n\,\mathbf{1}_\Lambda\big) \;=\; \lim_{n\to\infty} \mathbb{E}\big(\mathbb{E}(X \,|\, \mathcal{G}) \cdot Y_n\,\mathbf{1}_\Lambda\big) \;=\; \mathbb{E}\big(\mathbb{E}(X \,|\, \mathcal{G}) \cdot Y\,\mathbf{1}_\Lambda\big)$$

for every $\Lambda \in \mathcal{G}$. Thus (8.13) holds when X and Y are both nonnegative; for the general case, decompose each of these random variables in its positive and negative parts, and use the linearity property (8.18) below. □

Proposition 8.4. Independence makes conditioning irrelevant: *If the random variable X is integrable and the σ-algebras $\sigma(X)$ and $\mathcal{G}$ are independent, then*

$$\mathbb{E}(X \,|\, \mathcal{G}) \;=\; \mathbb{E}(X)\,, \quad \text{a.s.} \tag{8.15}$$

This is the case, in particular, if $\mathcal{G} = \{\emptyset, \Omega\}$ is the trivial $\sigma-$algebra.

More generally, if $\mathcal{H}$ is another sub-σ-algebra of $\mathcal{F}$ such that $\mathcal{G}$ and $\sigma\big(\sigma(X), \mathcal{H}\big)$ are independent, then

$$\mathbb{E}\,\big(X \,|\, \mathcal{G} \vee \mathcal{H}\big) \;=\; \mathbb{E}(X \,|\, \mathcal{H})\,, \qquad a.\text{s.} \tag{8.16}$$

where $\mathcal{G} \vee \mathcal{H} \equiv \sigma\big(\mathcal{G}, \mathcal{H}\big)$ is the σ-algebra generated by the family of sets

$$\mathcal{J} \;:=\; \big\{A \in \mathcal{F} \,|\, A = G \cap H \;\text{ with }\; G \in \mathcal{G}\,,\; H \in \mathcal{H}\,\big\}\,.$$

Proof: Because the random variables X and $\mathbf{1}_\Lambda$ are independent for $\Lambda \in \mathcal{G}$, Theorem 7.1 gives $\mathbb{E}(X \cdot \mathbf{1}_\Lambda) \;=\; \mathbb{E}(X) \cdot \mathbb{P}(\Lambda) \;=\; \mathbb{E}(\mathbb{E}(X) \cdot \mathbf{1}_\Lambda)$. This settles the first claim.

For the second, observe that the family of sets $\mathcal{J}$ is a π-system: closed under finite intersections. Assuming first that X is nonnegative and denoting by Y a version of the conditional expectation $\mathbb{E}(X \,|\, \mathcal{H})$, we have

$$\mathbb{E}\,(X \cdot \mathbf{1}_A) \;=\; \mathbb{E}\,(X\,1_H \cdot \mathbf{1}_G) \;=\; \mathbb{P}(G) \cdot \mathbb{E}(X\,\mathbf{1}_H)$$

and

$$\mathbb{E}(Y \cdot \mathbf{1}_A) \;=\; \mathbb{E}\,\big(Y\,\mathbf{1}_H \cdot \mathbf{1}_G\big) \;=\; \mathbb{P}(G) \cdot \mathbb{E}\big(Y\,\mathbf{1}_H\big) \;=\; \mathbb{P}(G) \cdot \mathbb{E}\big(X\,\mathbf{1}_H\big)$$

for every $G \in \mathcal{G}$, $H \in \mathcal{H}$, and with $A := G \cap H$. Now the two measures $\mu(A) := \mathbb{E}(X\,\mathbf{1}_A)$, $\nu(A) := \mathbb{E}(Y\,\mathbf{1}_A)$ have the same total mass

$$\nu(\Omega) \;=\; \mathbb{E}(Y) \;=\; \mathbb{E}\big(\mathbb{E}(X \,|\, \mathcal{H})\big) \;=\; \mathbb{E}(X) \;=\; \mu(\Omega) < \infty$$

and we have just shown that they agree on the π-system $\mathcal{J}$. From Theorem 4.8 and its Corollary, they agree also on $\mathcal{G} \vee \mathcal{H} \equiv \sigma(\mathcal{J})$, so $\mathbb{E}(X \cdot \mathbf{1}_A) = \mathbb{E}(Y \cdot \mathbf{1}_A)$ for all $A \in \mathcal{G} \vee \mathcal{H}$, thus proving (8.16). □

Example. Let $X, X_1, X_2, \ldots$ be independent random variables with common distribution $\mu = \mathbb{P} \circ X^{-1}$, $\mathbb{E}(|X|) < \infty$, and set $S_n := \sum_{j=1}^n X_j$ for $n \geq 1$. With the notation

$$\mathcal{G}_n \,:=\, \sigma\big(S_n, S_{n+1}, S_{n+2}, \ldots\big) \,=\, \sigma\big(S_n, X_{n+1}, X_{n+2}, \ldots\big) = \sigma\big(S_n\big) \vee \sigma\big(X_{n+1}, X_{n+2}, \ldots\big)\,,$$

we claim that for $n \geq 2$ we have:

$$\overline{X}_n \,:=\, \frac{S_n}{n} \,=\, \mathbb{E}\left(\, X_1 \,|\, \mathcal{G}_n \right), \quad \text{a.e.} \tag{8.17}$$

To see this, let us start by observing that $\sigma\big(X_{n+1}, X_{n+2}, \ldots\big)$ is independent of $\sigma\big(X_1, S_n\big)$, so (8.16) gives $\mathbb{E}(X_1 \,|\, \mathcal{G}_n) \,=\, \mathbb{E}(X_1 \,|\, \sigma(S_n)) \,=\, \mathbb{E}(X_1 \,|\, S_n)$, a.e. However, we have

$$\begin{aligned} \mathbb{E}\big(X_1 \cdot \mathbf{1}_{\{S_n \in B\}}\big) \,&=\, \int \ldots \int_{\{(x_1 + \ldots + x_n) \in B\}} x_1 \, \mathrm{d}\mu(x_1) \ldots \mathrm{d}\mu(x_n) \\ &= \mathbb{E}\big(X_2 \cdot \mathbf{1}_{\{S_n \in B\}}\big) \,=\, \ldots = \mathbb{E}\big(X_n \cdot \mathbf{1}_{\{S_n \in B\}}\big)\,, \quad \forall\, B \in \mathcal{B}(\mathbb{R})\,, \end{aligned}$$

and thanks to (8.13) and the linearity property (8.18) below:

$$\mathbb{E}\big(X_1 \,|\, \mathcal{G}_n\big) \,=\, \mathbb{E}(X_1 \,|\, S_n) \,= \ldots = \mathbb{E}\big(X_n \,|\, S_n\big) \,=\, \frac{1}{n} \cdot \mathbb{E}\big(X_1 + \ldots + X_n \,|\, S_n\big) \,=\, \frac{S_n}{n}\,. \qquad \square$$

Conditional Expectations (C.E.'s for short) obey the familiar rules of integration established in Chapter 4. In particular, there are the following analogues of the classical results, all valid for **integrable** random variables X and $\{X_n\}_{n \in \mathbb{N}}$. The proofs are straightforward; we supply only two, leaving the rest as exercises for the reader.

• **LINEARITY:** *For any real numbers α, β we have*

$$\mathbb{E}(\alpha X_1 + \beta X_2 \,|\, \mathcal{G}) \,=\, \alpha \cdot \mathbb{E}(X_1 \,|\, \mathcal{G}) \,+\, \beta \cdot \mathbb{E}(X_2 \,|\, \mathcal{G})\,, \quad \text{a.e.} \tag{8.18}$$

• **MONOTONICITY OF C.E.'s:** *If $X_1 \leq X_2$, then we have*

$$\mathbb{E}(X_1 | \, \mathcal{G}) \,\leq\, \mathbb{E}(X_2 | \, \mathcal{G})\,, \quad \text{a.e.} \tag{8.19}$$

• **MONOTONE CONVERGENCE THEOREM FOR C.E.'s:** *If we have $0 \leq X_1 \leq X_2 \leq \cdots$ and $X := \lim_{n \to \infty} \uparrow X_n$, then*

$$\lim_{n \to \infty} \uparrow\, \mathbb{E}\,(X_n | \, \mathcal{G}) \,=\, \mathbb{E}(X | \, \mathcal{G})\,, \quad \text{a.e.} \tag{8.20}$$

• **FATOU'S LEMMA FOR C.E.'s:** *If $X_n \geq 0$ for all $n \in \mathbb{N}$, then we have*

$$\mathbb{E}\left(\liminf_{n \to \infty} X_n | \, \mathcal{G}\right) \,\leq\, \liminf_{n \to \infty} \mathbb{E}(X_n | \, \mathcal{G}\,)\,, \quad \text{a.e.} \tag{8.21}$$

• **DOMINATED CONVERGENCE THEOREM FOR C.E.'s:** *If* $|X_n| \leq Y$ *for all* $n \in \mathbb{N}$ *and some integrable random variable* Y*, and if* $\lim_{n\to\infty} X_n = X$ *holds a.e., then we have*

$$\lim_{n\to\infty} \mathbb{E}\,(X_n \,|\, \mathcal{G}\,) \;=\; \mathbb{E}(X \,|\, \mathcal{G}\,)\,, \quad \text{a.e.} \tag{8.22}$$

• **CONDITIONAL JENSEN INEQUALITY:** *If* $\varphi : \mathbb{R} \to \mathbb{R}$ *is convex and the random variable* $\varphi(X)$ *is integrable, then we have*

$$\varphi\big(\,\mathbb{E}\big(X|\,\mathcal{G}\,)\,\big) \;\leq\; \mathbb{E}\,\big(\varphi(X)\,|\,\mathcal{G}\,\big)\,, \quad \text{a.e.} \tag{8.23}$$

• **CONDITIONING DECREASES** $\mathbb{L}^p$ **NORMS:** *For any* $p \in [1,\infty]$*, we have*

$$\Big|\Big|\,\mathbb{E}(\,X|\,\mathcal{G}\,)\,\Big|\Big|_p \;\leq\; ||X||_p\;. \tag{8.24}$$

Proof of (8.19): Take $X_1 \equiv 0$ and write $X \equiv X_2$, $Y \equiv \mathbb{E}(X\,|\,\mathcal{G})$. If $\mathbb{P}(Y \geq 0) = 1$ did not hold, then the set $G_n := \{Y < -1/n\} \in \mathcal{G}$ would have positive probability for some $n \in \mathbb{N}$, thus leading to the absurdity

$$0 \;\leq\; \int_{G_n} X\,\mathrm{d}\mathbb{P} \;=\; \int_{G_n} Y\,\mathrm{d}\mathbb{P} \;\leq\; -\frac{1}{n}\,\mathbb{P}(G_n) \;<\; 0\,.$$

Proof of (8.20): If Y_n is a version of the conditional expectation $\mathbb{E}(X_n|\,\mathcal{G}\,)$, then $Y_n \geq 0$ and $Y_n \leq Y_{n+1}$ hold a.e. (for all $n \in \mathbb{N}$). Thus, $Y := \lim_{n\to\infty} \uparrow Y_n$ exists and is nonnegative, $\mathcal{G}-$measurable. But now we can use the Monotone Convergence Theorem, letting $n \to \infty$ in $\mathbb{E}(X_n \cdot \mathbf{1}_\Lambda) \;=\; \mathbb{E}(Y_n \cdot \chi_\Lambda)$ to deduce $\mathbb{E}(X \cdot \mathbf{1}_\Lambda) \;=\; \mathbb{E}(Y \cdot \mathbf{1}_\Lambda)$, for all $\Lambda \in \mathcal{G}$. In other words, Y is a version of the conditional expectation $\mathbb{E}(X\,|\,\mathcal{G}\,)$.

Conditional expectations have also excellent continuity properties relative to monotone sequences of $\sigma-$algebras. We state below the most basic relevant result; its proof is best given in the context of the theory of Martingales, where it becomes a relatively easy exercise.

Proposition 8.5. P. LÉVY Convergence: *Suppose* Z *is an integrable random variable, and* $\big(\mathcal{G}_n\big)_{n\in\mathbb{N}}$ *an increasing (respectively, decreasing) sequence of* $\sigma-$*algebras, for which we define* $\mathcal{G}_\infty := \sigma\big(\bigcup_{n\in\mathbb{N}} \mathcal{G}_n\big)$ *(respectively,* $\mathcal{G}_\infty := \bigcap_{n\in\mathbb{N}} \mathcal{G}_n$*). Then*

$$\lim_{n\to\infty} \mathbb{E}\big(Z\,\big|\,\mathcal{G}_n\big) \;=\; \mathbb{E}\big(Z\,\big|\,\mathcal{G}_\infty\big)\,, \quad \mathbb{P}-\textit{a.e. and in}\ \mathbb{L}^1.$$

Proposition 8.6. BOREL Measurability: *Suppose that* $\mathcal{G} = \sigma(\Xi)$ *for some random variable* $\Xi : \Omega \to \mathbb{R}$*, and that the random variable* $X : \Omega \to \mathbb{R}$ *is integrable. Then there exists a Borel-measurable function* $h : \mathbb{R} \to \mathbb{R}$ *such that*

$$\mathbb{E}(X\,|\,\mathcal{G}\,) \;\equiv\; \mathbb{E}(X\,|\,\Xi\,) \;=\; h(\Xi)\,, \quad \mathbb{P}-\textit{a.e.}$$

And if we denote by $\mu = \mathbb{P} \circ \Xi^{-1}$ *the distribution of the random variable* Ξ *and set*

$$\nu(B) \,:=\, \int_{\Xi^{-1}(B)} X \,\mathrm{d}\mathbb{P} \,=\, \mathbb{E}(X\,\mathbf{1}_B(\Xi))\,, \qquad B \in \mathcal{B}(\mathbb{R})\,,$$

then this function h *is a version of the* RADON-NIKODÝM *derivative of* ν *with respect to* μ*, that is:* $h = \mathrm{d}\nu/\mathrm{d}\mu$*,* μ*-a.e. on* $\mathbb{R}$*; equivalently, for every* $B \in \mathcal{B}(\mathbb{R})$ *we have*

$$\mathbb{E}\big(X\,\mathbf{1}_B(\Xi)\big) \,=\, \int_B h\,d\mu\,.$$

Proof: The first claim follows from Exercise 2.2. For the second, take arbitrary $B \in \mathcal{B}(\mathbb{R})$ and note that the quantity

$$\nu(B) \,=\, \int_{\Xi^{-1}(B)} X \,\mathrm{d}\mathbb{P} \,=\, \int_{\Xi^{-1}(B)} \mathbb{E}(X \,|\, \Xi\,) \,\mathrm{d}\mathbb{P} \,=\, \int_{\Xi^{-1}(B)} h(\Xi\,) \,\mathrm{d}\mathbb{P}$$

is equal to

$$\mathbb{E}\big(\,h(\Xi\,)\,\mathbf{1}_B(\Xi)\,\big) \,=\, \int_{\mathbb{R}} h\,\mathbf{1}_B \,\mathrm{d}\mu \,=\, \int_B h\,\mathrm{d}\mu\,. \qquad\qquad \square$$

Remark 8.4. Abusing notation slightly, we shall denote the function of Proposition 8.6 by

$$h(\xi) \,=\, \mathbb{E}(X \,|\, \Xi = \xi)\,, \qquad \xi \in \mathbb{R}$$

and write, for instance,

$$\mathbb{E}\big(X\,\mathbf{1}_B(\Xi)\big) \,=\, \int_B \mathbb{E}(\,X \,|\, \Xi = \xi\,) \,\mathrm{d}\mu(\xi)\,, \qquad B \in \mathcal{B}(\mathbb{R})\,.$$

(It is important, however, always to bear in mind that this function h is defined only for $\mu-$a.e. $\xi \in \mathbb{R}$, and to be very careful before making bold statements about it!)

All this generalizes in a natural way to a finite number $\Xi_1, \cdots, \Xi_n$ of random variables: there exists a BOREL-measurable function $h : \mathbb{R}^n \to \mathbb{R}$ such that $\mathbb{E}(X \,|\, \Xi_1, \cdots, \Xi_n\,) \,=\, h(\Xi_1, \cdots, \Xi_n)$ holds $\mathbb{P}-$a.e., and we write $\mathbb{E}(X \,|\, \Xi_1 = \xi_1, \cdots, \Xi_n = \xi_n)$ for $h(\xi_1, \cdots, \xi_n)$.

Remark 8.5. Let us make contact now with the more classical notions of conditional expectations and probabilities that we encounter, say, when dealing with a vector (Y, Ξ) of random variables that have joint distribution

$$\mathbb{P}[\,(Y, \Xi) \in A\,] \,=\, \int \int_A f_{Y,\Xi}(y, \xi) \,\mathrm{d}y \,\mathrm{d}\xi\,, \qquad \forall \quad A \in \mathcal{B}(\mathbb{R}^2)$$

for some probability density function $f_{Y,\Xi} : \mathbb{R}^2 \to [0, \infty)$. Let us set $f_\Xi(\xi) := \int_{\mathbb{R}} f_{Y,\Xi}(y, \xi)\,dy$ for the 'marginal probability density function' for Ξ;

$$f_{Y|\Xi}(y|\xi) \,:=\, \frac{f_{Y,\Xi}(y, \xi)}{f_\Xi(\xi)}\,, \quad y \in \mathbb{R} \quad \text{ when } \; f_\Xi(\xi) > 0$$

and $f_{Y|\Xi}(\,\cdot\,|\xi) := 0$ when $f_\Xi(\xi) = 0$, for the 'conditional probability density of Y given $\{\Xi = \xi\}$' ; as well as

$$h(\xi) := \int_{\mathbb{R}} g(y)\, f_{Y|\Xi}(y|\xi)\, \mathrm{d}y\,, \quad \xi \in \mathbb{R}\,,$$

where $g : \mathbb{R} \to [0,\infty)$ is any given Borel-measurable function $g : \mathbb{R} \to [0,\infty)$. If $\mathbb{E}(g(Y)) < \infty$, then we claim that

$$h(\,\Xi\,) \text{ **is a version of the conditional expectation** } \mathbb{E}(\,g(Y)\,|\,\Xi\,)\,.$$

Indeed, we have to show

$$\mathbb{E}[\,g(Y)\,\mathbf{1}_B(\Xi)\,] \;=\; \mathbb{E}[\,h(\Xi)\,\mathbf{1}_B(\Xi)\,]$$

for every Borel set B; but the left-hand-side of this expression is given by the double integral

$$\int\int_{\mathbb{R}^2} g(y)\mathbf{1}_B(\xi) f_{Y,\Xi}(y,\xi)\, dyd\xi\,,$$

whereas TONELLI's theorem allows us to express the right-hand-side as

$$\int_B h(\xi)\, f_\Xi(\xi)\, \mathrm{d}\xi \;=\; \int\int_{\mathbb{R}^2} g(y)\mathbf{1}_B(\xi)\, f_{Y|\Xi}(y|\xi)\, f_\Xi(\xi)\, \mathrm{d}y\, \mathrm{d}\xi \;=\; \int\int_{\mathbb{R}^2} g(y)\mathbf{1}_B(\xi)\, f_{Y,\Xi}(y,\xi)\, \mathrm{d}y\, \mathrm{d}\xi\,.$$

This computation gives meaning to the intuitive statement, that $h(\xi)$ is the *"conditional expectation of $g(Y)$, given $\Xi = \xi$"*. □

Exercise 8.3. Let $X_1, X_2, \cdots$ be a sequence of independent, strictly positive random variables with the same, nondegenerate distribution function $F(x) = \mathbb{P}(X_k \le x)$, $x \in \mathbb{R}$ for all $k \in \mathbb{N}$ that satisfies $F(0) = 1$. Suppose the variable X_n represents a potential reward which is available to you on day $t = n$, should you decide to terminate the game on that day, collect your reward, and go home. Suppose you adopt the following rule: "wait until the first day $t = n$ whose reward X_n is strictly bigger than the first reward X_1 you have observed"; that is, you terminate the game at the random time

$$T \;=\; \inf\big\{\, n \ge 2 \,|\, X_n > X_1 \,\big\}$$

if the indicated set is nonempty, otherwise you are damned to keep playing forever ($T = \infty$) and collect nothing.

What is the distribution of T? of your actual reward $X_T\,\mathbf{1}_{\{T<\infty\}}$? What is the expected duration of the game? your expected reward? how does your expected reward compare to $\mathbb{E}(X_1)$, the expected reward you would have collected, had you stopped the game on the first day?

Exercise 8.4. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a complete probability space, and let $\mathcal{G}$, $\mathcal{H}$, $\mathcal{U}$ be sub$-\sigma-$ algebras of $\mathcal{F}$ such that

$$\mathcal{G} \subseteq \mathcal{H}\,, \quad \mathcal{H} \text{ is independent of } \mathcal{U}\,, \quad \mathcal{G} \vee \mathcal{U} \;=\; \mathcal{H} \vee \mathcal{U}\,.$$

Show that we have then $\mathcal{G} = \mathcal{H}$.

Exercise 8.5. RADON-NIKODÝM Derivatives: Let $\mathbb{P}$, $\mathbb{Q}$ be probability measures on $(\Omega, \mathcal{F})$ with $\mathbb{Q} < \mathbb{P}$ and denote by

$$X := \frac{d\mathbb{Q}}{d\mathbb{P}}\Big|_{\mathcal{F}}, \qquad Z := \frac{d\mathbb{Q}}{d\mathbb{P}}\Big|_{\mathcal{G}}$$

the Radon-Nikodým derivatives of $\mathbb{Q}$ with respect to $\mathbb{P}$ on $\mathcal{F}$ and on $\mathcal{G}$, respectively, for some sub-σ-algebra $\mathcal{G}$ of $\mathcal{F}$. Show that, with $\mathbb{E}^{\mathbb{P}}[\,\cdot\,]$ denoting expectation with respect to $\mathbb{P}$, we have

$$Z \;=\; \mathbb{E}^{\mathbb{P}}\big(X \,|\, \mathcal{G}\big)\,, \quad \text{a.s.} \tag{8.25}$$

Exercise 8.6. Conditioning Decreases the Relative Entropy: In the setting of Exercise 8.5, let us denote by

$$H_{\mathcal{F}}(\mathbb{Q}\,|\,\mathbb{P}) \;:=\; \mathbb{E}^{\mathbb{Q}}\big(\log X\big) \;=\; \mathbb{E}^{\mathbb{P}}\big(X \log X\big)\,, \qquad H_{\mathcal{G}}(\mathbb{Q}\,|\mathbb{P}) \;:=\; \mathbb{E}^{\mathbb{Q}}\big(\log Z\big) \;=\; \mathbb{E}^{\mathbb{P}}\big(Z \log Z\big)$$

the relative entropy of $\mathbb{Q}$ with respect to $\mathbb{P}$ on the σ-algebras $\mathcal{F}$ and $\mathcal{G}$, respectively. Show that we have

$$H_{\mathcal{F}}(\mathbb{Q}\,|\,\mathbb{P}) \;\geq\; H_{\mathcal{G}}(\mathbb{Q}\,|\,\mathbb{P})\,. \tag{8.26}$$

Then use this property to establish the PINSKER-CSISZÁR *inequality*

$$H_{\mathcal{F}}(\mathbb{Q}\,|\,\mathbb{P}) \;\geq\; 2 \cdot \Big(||\mathbb{Q} - \mathbb{P}||_{\mathcal{F}}\Big)^2\,, \tag{8.27}$$

where $||\mathbb{Q} - \mathbb{P}||_{\mathcal{F}} := \sup_{A \in \mathcal{F}} |\mathbb{Q}(A) - \mathbb{P}(A)|$ is the total variation distance of the two measures $\mathbb{Q}$ and $\mathbb{P}$ on $\mathcal{F}$.

(*Hint:* For (8.26), observe that the function $f(x) = x \log x$, $x > 0$ is convex. For (8.27), start by establishing the elementary inequality

$$q \cdot \log(q/p) + (1-q) \cdot \log\big((1-q)/(1-p)\big) \geq 2(p-q)^2$$

for $0 < p,\, q < 1$.)

Exercise 8.7. On a complete probability space $(\Omega, \mathcal{F}, \mathbb{P})$ consider integrable random variables $X_1, \cdots, X_n$ and $X_{n+1} = X_1$, as well as sub-σ-algebras $\mathcal{F}_1, \cdots, \mathcal{F}_n$ of $\mathcal{F}$, for which

$$\mathbb{E}\big(X_{i+1} \,|\, \mathcal{F}_i\big) \;=\; X_i\,, \quad \mathbb{P}-\text{a.s.}$$

holds for every $i = 1, \cdots n$. Show that $X_1 = \cdots = X_n$ holds a.e.

(*Hint:* Assume first that the X_i's are square-integrable; then truncate.)

Exercise 8.8. Let $X_1, X_2, \cdots$ be independent random variables with respective distributions $\mu_j = \mathbb{P} \circ X_j^{-1}$. Then for any BOREL set B of the real line, we have

$$\mathbb{P}(\, X_1 + X_2 \in B \,|\, X_1\,) \;=\; \mu_2(B - x)\big|_{x = X_1}\,, \qquad \text{a.s.;} \tag{8.28}$$

and more generally, with $S_n = \sum_{j=1}^n X_j$, we have:

$$\mathbb{P}(\, S_n \in B \,|\, S_1, \cdots, S_{n-1}\,) \;=\; \mu_n(B - x)\big|_{x = S_{n-1}} \;=\; \mathbb{P}(\, S_n \in B \,|\, S_{n-1}\,)\,, \quad \text{a.s.} \tag{8.29}$$

8.1 Regular Conditional Probabilities

If $\{E_n\}_{n\in\mathbb{N}}$ is a sequence of disjoint sets in $\mathcal{F}$, then it is easily seen that

$$\mathbb{P}\Big(\bigcup_{n\in\mathbb{N}} E_n \,\Big|\, \mathcal{G}\Big)(\omega) \;=\; \sum_{n\in\mathbb{N}} \mathbb{P}(E_n \,|\, \mathcal{G})(\omega) \qquad \text{holds for a.e. } \omega\in\Omega\,. \tag{8.30}$$

This does *not* imply, however, that for every (or even almost every) $\omega\in\Omega$, the set function

$$\mathcal{F}\ni E \;\longmapsto\; \mathbb{P}(E\,|\,\mathcal{G})(\omega) \;\in\; [0,1] \qquad \text{is a measure.} \tag{8.31}$$

Trouble can arise from the fact that the exceptional set in (8.30) may very well depend on the sequence $\{E_n\}_{n\in\mathbb{N}}$ itself; thus the union of such exceptional sets, over all possible sequences of disjoint sets $\{E_n\}_{n\in\mathbb{N}}\subset\mathcal{F}$, may become quite large – indeed cover Ω itself!

In many contexts it becomes therefore desirable, often necessary, to have conditions guaranteeing (8.31) in some form. This is the subject of the discussion that follows.

Definition 8.3. Regular Conditional Probabilities: Let $X:\Omega\to S$ be a measurable mapping from the probability space $(\Omega,\mathcal{F},\mathbb{P})$ into the measurable space $(S,\mathcal{S})$, and consider a sub-σ-algebra $\mathcal{G}$ of $\mathcal{F}$. A *regular conditional probability for X given $\mathcal{G}$*, is a mapping $Q:\Omega\times\mathcal{S}\to[0,1]$ with the following properties:

(i) $B\mapsto Q(\omega,B)$ is a probability measure on $(S,\mathcal{S})$, for all $\omega\in\Omega$;

(ii) for every $B\in\mathcal{S}$, the mapping $\omega\mapsto Q(\omega,B)$ is $\mathcal{G}$-measurable; and

(iii) for every $B\in\mathcal{S}$, we have

$$\mathbb{P}[\,X\in B\,|\,\mathcal{G}\,](\omega) \;=\; Q(\omega,B) \qquad \text{for a.e. } \omega\in\Omega\,. \qquad\qquad \diamond$$

We refer to PARTHASARATHY (1967), pp. 146-150, for the following fundamental result.

Theorem 8.1. Existence of Regular Conditional Probabilities: *In the setting of Definition 8.3, a regular conditional probability for X given $\mathcal{G}$ exists, provided that S is a complete, separable metric space and $\mathcal{S}=\mathcal{B}(S)$ is the σ-algebra of its* BOREL *sets.*

9 Simple Random Walk

We shall study in this chapter the sequence of random variables

$$S_n \;=\; s + \sum_{j=1}^{n} X_j \,, \qquad n \in \mathbb{N}_0 \tag{9.1}$$

with $S_0 = s \in \mathbb{R}$ a given "starting position" and $X_1,\, X_2, \cdots$ a sequence of independent BERNOULLI random variables with

$$\mathbb{P}(X_j = +1) \;=\; p \,, \quad \mathbb{P}(X_j = -1) \;=\; 1 - p =: q \tag{9.2}$$

for some given $p \in (0,1)$. The resulting sequence of random variables $\{S_n\}_{n\in\mathbb{N}_0}$ is called *Simple Random Walk* with starting position s. In the special case $p = 1/2$, we shall refer to this sequence as the *Simple, Symmetric Random Walk*.

We shall make a habit, to incorporate the starting position as a subscript to the underlying probability measure, so we shall be writing generically $\mathbb{P}_s$ rather than simply $\mathbb{P}$ from now on. Let us observe that with

$$A_n \;:=\; \sum_{j=1}^{n} \mathbf{1}_{\{X_j=+1\}} \,, \qquad B_n \;:=\; \sum_{j=1}^{n} \mathbf{1}_{\{X_j=-1\}}$$

denoting, respectively, the number of "up" and of "down" moves that the particle has made up to time $t = n$, we have $A_n + B_n = n$ and $A_n - B_n = S_n - s$, so that $A_n = \big(S_n - s + n\big)/2$ has $Bin\,(n;p)$ distribution:

$$\mathbb{P}_s\big(S_n - s = x\big) \;=\; N_n(x)\; p^{\,(n+x)/2}\, q^{\,(n-x)/2} \;=:\; \mathfrak{b}_n(x) \,, \qquad x \in \mathbb{Z} \,, \tag{9.3}$$

with the interpretation $\mathfrak{b}_n(x) = 0$ for $n + x < 0$ or $n + x > 2n$ and with the notation

$$N_n(x) \;:=\; \binom{n}{(n+x)/2} \,. \tag{9.4}$$

9.1 Always Ahead, Never Behind

We begin with the following question: *Suppose that a match of ping-pong ended 21:18. What is the probability that the victor was strictly ahead throughout the entire match ?* Assuming bravely that the outcomes of the difference subgames of the match are independent, and the eventual victor had the same, constant probability of winning any and each one of them, the following result provides the answer $1/13$, roughly 7.7 %.

Theorem 9.1. Always Ahead: *For any $n \in \mathbb{N}$ and $x = 1, \cdots n$ we have*

$$\boxed{\mathbb{P}_0\big(S_1 > 0\,,\, \cdots\,,\, S_{n-1} > 0 \,|\, S_n = x\big) \;=\; \frac{x}{n} \,.} \tag{9.5}$$

Proof: In view of (9.3), we need to compute $\mathbb{P}_0\big(S_1 > 0\,,\,\cdots\,,\,S_{n-1} > 0\,,\,S_n = x\big)$; this is equal to $p^{(n+x)/2}\,q^{(n-x)/2}$, times

$$\begin{aligned}
&\#\,\big(\text{paths from (0,0) to } (n,x) \text{ strictly above the axis}\big) \\
&= \#\,\big(\text{paths from (1,1) to } (n,x) \text{ strictly above the axis}\big) \\
&= \#\,\big(\text{paths from (1,1) to } (n,x)\big) - \#\,\big(\text{paths from (1,1) to } (n,x) \text{ that touch the axis}\big) \\
&= \#\,\big(\text{paths from (0,0) to } (n-1,x-1)\big) - \#\,\big(\text{paths from (1,1) to } (n,x) \text{ that touch the axis}\big).
\end{aligned}$$

In the notation of (9.4), we have the first term in this last expression as

$$\#\,\big(\text{paths from (0,0) to } (n-1,x-1)\big) \;=\; N_{n-1}(x-1)\,.$$

As for the second term, we obtain it from the following, renowned, and completely obvious by just drawing a picture, *Reflection Principle* of DÉSIRÉ ANDRÉ:

$$\#\,\big(\text{paths from (1,1) to } (n,x) \text{ that touch the axis}\big) \;=\; \#\,\big(\text{paths from } (1,-1) \text{ to } (n,x)\big)\,,$$

and by a "northwest shift", that is, a retreat by one step in the first coördinate and an advance by one step in the second:

$$\#\,\big(\text{paths from } (1,-1) \text{ to } (n,x)\big) \;=\; \#\,\big(\text{paths from } (0,0) \text{ to } (n-1,x+1)\big) \;=\; N_{n-1}(x+1)\,.$$

Putting all this together, we obtain after some mild algebraic manipulation

$$\begin{aligned}
&\#\,\big(\text{paths from (0,0) to } (n,x) \text{ strictly above the axis}\big) \;= \\
&\qquad = \; N_{n-1}(x-1) - N_{n-1}(x+1) \;=\; \frac{x}{n}\,N_n(x)
\end{aligned} \tag{9.6}$$

(see also below), as well as the formula

$$\boxed{\mathbb{P}_0\big(S_1 > 0\,,\,\cdots\,,\,S_{n-1} > 0\,,\,S_n = x\big) \;=\; p^{(n+x)/2}\,q^{(n-x)/2}\,\cdot\,\big(N_{n-1}(x-1) - N_{n-1}(x+1)\big)} \tag{9.7}$$

which is quite important in its own right.

Comparing this expression with (9.3), we obtain

$$\begin{aligned}
\mathbb{P}_0\big(S_1 > 0\,,\,\cdots\,,\,S_{n-1} > 0\,\big|\,S_n = x\big) \;&=\; \frac{N_{n-1}(x-1) - N_{n-1}(x+1)}{N_n(x)} \\
&= \; \frac{\binom{n-1}{a-1} - \binom{n-1}{a}}{\binom{n}{a}} \;=\; \frac{a-b}{n} \;=\; \frac{x}{n}\,,
\end{aligned}$$

where we have set $a = (n+x)/2\,,\; b = (n-x)/2$. The expression of (9.5) follows. □

Let us recall now the question we asked about the 21:18 ping-pong match right before Theorem 9.1, but suppose we choose to pose a slightly different one: *What is the probability that the victor never fell behind throughout the match?* Under the same assumptions as before, the following result provides the answer $2/11 \approx 18.2\%$, a number significantly bigger than before.

Theorem 9.2. Never Behind: *For any* $n \in \mathbb{N}$ *and* $x = 1, \cdots n$ *we have*

$$\boxed{\mathbb{P}_0\big(S_1 \geq 0\,,\,\cdots\,,\,S_{n-1} \geq 0 \mid S_n = x\big) \;=\; \frac{x+1}{\big((n+x)/2\big)+1}\;.} \tag{9.8}$$

Proof: Just as before, we compute $\mathbb{P}_0\big(S_1 \geq 0\,,\,\cdots\,,\,S_{n-1} \geq 0\,,\,S_n = x\big)$ as $p^{\,(n+x)/2}\,q^{\,(n-x)/2}$, times

$$\begin{aligned} &\#\,\big(\text{paths from (0,0) to } (n,x) \text{ that do not fall below the axis}\big) \;= \\ &\quad =\; \#\,\big(\text{paths from (-1,-1) to } (n,x) \text{ that stay strictly above the level } -1\big) \end{aligned}$$

$$=\; \#\,\big(\text{paths from (0,0) to } (n+1,x+1) \text{ that stay strictly above the axis}\big) \;=\; \frac{x+1}{n+1}\,N_{n+1}(x+1)$$

from (9.6) and (9.4); again, it is very helpful to draw a picture. The expression of (9.8) follows readily from this and (9.3). □

9.2 First Passages

It is not so hard actually to build on this theorem, and obtain some rather deep results about the behavior of the simple, symmetric random walk. In order to make headway let us introduce the *first hitting time* and the *first passage time*

$$H_k \;:=\; \inf\{\,n \in \mathbb{N}_0 \,|\, S_n = k\,\}\,, \qquad T_k \;:=\; \inf\{\,n \in \mathbb{N} \,|\, S_n = k\,\}\,, \tag{9.9}$$

respectively, for every element $k \in \mathbb{Z}$ of the integer lattice, the state-space of the simple random walk. The difference: with, say, $S_0 = 0$, on the event $\{X_1 = +1\,,\,X_2 = -1\}$ we have $H_0 = 0$ but $T_0 = 2$.

Theorem 9.3. Null Recurrence of the Simple, Symmetric Random Walk: *Consider the simple, symmetric random walk of (9.1) with* $p = 1/2$. *For any* $n \in \mathbb{N}$ *we have*

$$2\,\mathbb{P}_0\big(\,S_1 > 0\,,\,\cdots\,,\,S_{2n-1} > 0\,,\,S_{2n} > 0\,\big) \;=\; \mathbb{P}_0\big(S_{2n} = 0\big) \;=\; \binom{2n}{n}\left(\frac{1}{2}\right)^{2n}\,. \tag{9.10}$$

In particular, for the time T_0 *of first return to the origin, defined in (9.9), we have*

$$\boxed{\mathbb{P}_0\big(T_0 > 2n\big) \;=\; \mathbb{P}_0\big(S_{2n} = 0\big) \;=\; \binom{2n}{n}\left(\frac{1}{2}\right)^{2n} \;\sim\; \frac{1}{\sqrt{\pi\,n}}\,.} \tag{9.11}$$

It develops that we have

$$\mathbb{P}_0\big(T_0 = 2n\big) \;=\; \frac{1}{2n-1}\,\binom{2n-1}{n}\left(\frac{1}{2}\right)^{2n-1}\,, \quad n \in \mathbb{N} \tag{9.12}$$

as well as $\mathbb{P}_0(T_0 < \infty) = 1$, *but also* $\mathbb{E}_0(T_0) = \infty$.

Proof: Let us read (9.7) with $p = q = 1/2$:

$$\mathbb{P}_0\big(S_1 > 0\,,\,\cdots\,,\,S_{2n-1} > 0\,,\,S_{2n} = 2r\big) \;=\; \Big(N_{2n-1}(2r-1) - N_{2n-1}(2r+1)\Big)\cdot\left(\frac{1}{2}\right)^{2n}$$

$$= \;\frac{1}{2}\,\Big(\mathfrak{b}_{2n-1}(2r-1) - \mathfrak{b}_{2n-1}(2r+1)\Big)$$

in the notation of (9.3).

Summing this expression up over $r = 1, \cdots, n$ and exploiting the "telescoping" nature of the resulting series, we obtain

$$\mathbb{P}_0\big(S_1 > 0\,,\,\cdots\,,\,S_{2n-1} > 0\,,\,S_{2n} > 0\big) \;=\; \frac{1}{2}\,\mathfrak{b}_{2n-1}(1) \;=\; \frac{1}{2}\,\mathbb{P}_0\big(S_{2n} = 0\big)$$

(the last equality is just an easy computation), and we are done.

As for (9.11), it is rather obvious that $\mathbb{P}_0(T_0 > 2n)$ is twice $\mathbb{P}_0\big(S_1 > 0\,,\,\cdots\,,\,S_{2n} > 0\big)$, whereas the approximation for $n \to \infty$ follows from the STIRLING formula (3.7). The fact that this probability decays to zero shows that $\mathbb{P}_0(T_0 = \infty) = 0$, that is, T_0 is $\mathbb{P}_0-$a.s. finite, so the simple, symmetric random walk is recurrent – it returns, in fact infinitely often, to its starting point. But the fact that this decay is so slow that the series $\sum_{n\in\mathbb{N}} \mathbb{P}_0(T_0 > 2n)$ diverges, leads to the so-called "null recurrence" $\mathbb{E}_0(T_0) = \infty$. □

Exercise 9.1. In the context of this chapter and with the notation of (9.9), show

$$\mathbb{P}_0\big(T_k = k + 2n\big) \;=\; \frac{k}{k+2n}\,\binom{k+2n}{n}\,p^{\,k+n}\,q^{\,n}\,, \qquad n \in \mathbb{N}_0$$

for any given $k \in \mathbb{N}$. In particular, for $p = q = 1/2$ and $k = 1$ we have

$$\mathbb{P}_0\big(T_1 = 2n-1\big) \;=\; \frac{1}{2n-1}\,\binom{2n-1}{n}\left(\frac{1}{2}\right)^{2n-1} \;=\; \mathbb{P}_0\big(T_0 = 2n\big)\,, \quad n \in \mathbb{N}\,,$$

exactly the expression of (9.12).

Exercise 9.2. Complement the result of (9.10), by showing

$$\mathbb{P}_0\big(\,S_1 \geq 0\,,\,\cdots\,,\,S_{2n-1} \geq 0\,,\,S_{2n} \geq 0\,\big) \;=\; \mathbb{P}_0\big(S_{2n} = 0\big) \;=\; \binom{2n}{n}\left(\frac{1}{2}\right)^{2n}. \tag{9.13}$$

9.3 Last Visits

Let us fix now an integer N and consider, the last visit

$$L_{2N} \;:=\; \max\big\{\,0 \leq m \leq 2N \;\big|\; S_m = 0\,\big\}$$

to the origin by the random walk, before time $t = 2N$. What is the distribution of this random variable? It turns out we have now all the tools to reduce this question to an (almost) triviality.

Theorem 9.4. Discrete Arc-Sine Distribution: *For the simple symmetric random walk the distribution of the last visit to zero is given by*

$$\boxed{\mathbb{P}_0\big(L_{2N}=2n\big) \;=\; \binom{2n}{n}\binom{2N-2n}{N-n}\left(\frac{1}{2}\right)^{2N} \quad , \qquad n=0,\cdots,N\,.}$$

Proof: Let us start with a simple but crucial observation; we shall uses it again and again in the study of random walk.

For any given $n \in \mathbb{N}_0$, we consider the sequence $\widehat{\mathfrak{S}}$ of random variables $\widehat{S}_j := S_{n+j} - S_n$, $j \in \mathbb{N}_0$. This is measurable with respect to $\sigma(X_{n+1}, X_{n+2}, \cdots)$, thus *independent of* $\sigma(X_1,\cdots,X_n) = \sigma(S_1,\cdots,S_n)$. *It is also a simple random walk with the same distribution as* $\mathfrak{S} = \{S_n\}_{n\in\mathbb{N}_0}$.

Keeping these considerations in mind, let us recall (9.11) and observe

$$\begin{aligned}\mathbb{P}_0\big(L_{2N}=2n\big) &= \mathbb{P}\big(S_{2n}=0, S_{2n+1}\neq 0,\cdots, S_{2N}\neq 0\big) = \mathbb{P}\big(S_{2n}=0, \widehat{S}_1\neq 0,\cdots,\widehat{S}_{2N}\neq 0\big)\\ &= \mathbb{P}\big(S_{2n}=0\big)\cdot\mathbb{P}\big(\widehat{S}_1\neq 0,\cdots,\widehat{S}_{2N}\neq 0\big) \;=\; \mathbb{P}\big(S_{2n}=0\big)\cdot\mathbb{P}\big(T_0>2(N-n)\big)\\ &= \mathbb{P}\big(S_{2n}=0\big)\cdot\mathbb{P}\big(S_{2(N-n)}=0\big) \;=\; \binom{2n}{n}\left(\frac{1}{2}\right)^{2n}\cdot\binom{2N-2n}{N-n}\left(\frac{1}{2}\right)^{2(N-n)}\,. \qquad \square\end{aligned}$$

By means of the STIRLING formula (3.7), we see that the expression of Theorem 9.4 can be approximated as

$$\mathbb{P}_0\big(L_{2N}=2n\big) \;\sim\; \frac{1}{\pi\sqrt{n(N-n)}}\,.$$

This way we write, for N large, and for $0\le x\le 1$:

$$\mathbb{P}_0\left(\frac{L_{2N}}{2N}\le x\right) \;=\; \sum_{\substack{n\in\mathbb{N}_0\\ n\le xN}}\mathbb{P}_0\big(L_{2N}=2n\big) \;\sim\; \sum_{\substack{n\in\mathbb{N}_0\\ n\le xN}}\frac{1}{\pi\sqrt{n(N-n)}}$$

$$\sim\; \sum_{\substack{n\in\mathbb{N}_0\\ (n/N)\le x}}\frac{1/N}{\pi\sqrt{(n/N)\big(1-(n/N)\big)}} \;\sim\; \int_0^x \frac{d\xi}{\pi\sqrt{\xi(1-\xi)}} \;=\; \frac{2}{\pi}\,\mathrm{Arcsin}\big(\sqrt{x}\,\big) =: F_{AS}(x)\,.$$

9.4 Gambler's Ruin

Suppose Paul has a proverbial *fistful of* $\$\,s$ *dollars*, and that Peter has another fistful $\$\,(b-s)$ for a total of $\$\,b$; we take $s<b$ to be natural numbers. The two friends keep tossing a coin: each time the toss turns up heads, Peter pays Paul \$1; the reverse for tails. The game ends when one of them ends up with all the capital in his hands. *Will this game end? how long (on the average) will it last? how likely is it that the eventual winner is Paul?*

Assuming that successive tosses $X_1,\, X_2,\, \cdots$ are independent with the common BERNOULLI distribution of (9.2), Paul's fortune S_n on day $t = n$ can be represented as in (9.1). Recalling the notation $H_k = \inf\{\, n \in \mathbb{N}_0 \,|\, S_n = k \,\}$ of (9.9), we would like to compute the probabilities

$$h(s) := \mathbb{P}_s(H_0 < H_b)\,, \qquad g(s) := \mathbb{P}_s(H_0 > H_b)$$

that Peter (respectively, Paul) wins the game, as well as the expected duration of the game

$$\mathfrak{d}(s) \;:=\; \mathbb{E}_s(D)\,, \qquad \text{where} \quad D \;:=\; H_0 \wedge H_b$$

for $s = 1, 2, \cdots, b-1$. Clearly, $h(0) = g(b) = 1$, $g(0) = h(b) = 0$ and $\mathfrak{d}(0) = \mathfrak{d}(b) = 0$.

We shall use *again and again* the familiar partition rule of section 7.3, by conditioning on the outcome X_1 of the first toss – along with the fact, already used in the proof of Theorem 9.4, that *the sequence* $\widehat{\mathfrak{S}}$ *of random variables* $\widehat{S}_n := S_{1+n} - S_1\,,\ n \in \mathbb{N}_0$ *is a simple random walk started at the origin and independent of the first toss* X_1.

In particular, these considerations give

$$h(s) \;=\; \mathbb{P}_s\big(H_0 < H_b \,|\, X_1 = 1\big)\, p + \mathbb{P}_s\big(H_0 < H_b \,|\, X_1 = -1\big)\, q \;=\; p\, h(s{+}1) + q\, h(s{-}1) \qquad (9.14)$$

for $s = 1, \cdots, b-1$. With $p \neq 1/2$, this difference equation $p\, h(s+1) - h(s) + q\, h(s-1) = 0$ has general solution $h(s) = A + B(q/p)^s$; the two constants are determined from the boundary conditions $h(0) = 1$, $h(b) = 0$ and we obtain

$$h(s) \;=\; \frac{\big(q/p\big)^s - \big(q/p\big)^b}{1 - \big(q/p\big)^b}\,, \qquad s = 0, 1, \cdots, b\,.$$

On the other hand, for $p = 1/2$ the difference equation of (9.14) has general solution $h(s) = A + Bs$ and now the boundary conditions $h(0) = 1$, $h(b) = 0$ lead to

$$h(s) \;=\; 1 - (s/b)\,, \qquad s = 0, 1, \cdots, b\,.$$

• In a completely analogous (and symmetric) fashion, one computes

$$g(s) \;=\; \frac{1 - \big(q/p\big)^s}{1 - \big(q/p\big)^b} \qquad \text{for } \ p \neq 1/2\,; \qquad g(s) \;=\; s/b \quad \text{for } \ p = 1/2$$

and verifies $h(s) + g(s) = 1$, for all $s = 0, 1, \cdots, b$. In particular, *the game ends with probability one* (please take a moment to argue this out).

• Suppose now that Peter is infinitely rich ($b = \infty$). Then Paul has no chance to win this game, so the question is how much of a chance $g^*(s) = \mathbb{P}_s(H_0 = \infty)$ he stands just to "stay alive", that is, not to see all his capital get wiped out in finite time.

We can compute formally the probability

$$h^*(s) \;=\; \mathbb{P}_s(H_0 < \infty) \;=\; \lim_{b \to \infty} \mathbb{P}(H_0 < H_b)$$

just by letting $b \to \infty$ in the above expression for $h(s)$, namely,

$$h^*(s) \,=\, \mathbb{P}_s(H_0 < \infty) \,=\, 1\,, \quad \text{for } \; 0 < p \le 1/2$$
$$= \big(q/p\big)^s\,, \quad \text{for } \; 1/2 < p < 1\,.$$

To put it a bit differently: *"the poor gets wiped out in the end, with probability one, even if the game is fair"*. [39] Only in a favorable game, i.e., one with $p > 1/2$, does the poor guy have a positive probability

$$g^*(s) \,=\, 1 - h^*(s) \,=\, \mathbb{P}_s(H_0 = \infty) \,=\, 1 - \big(q/p\big)^s\,, \qquad s = 1, 2, \cdots$$

to survive; and of course, the more his starting capital, the better his chances of eventual survival.

- For the expected duration $\mathfrak{d}(s) = \mathbb{E}_s(D)$ of the game, the partition rule gives

$$\mathfrak{d}(s) \,=\, \mathbb{E}_s\big(D \,|\, X_1 = 1\big)\, p + \mathbb{E}_s\big(D \,|\, X_1 = -1\big)\, q\,.$$

Now on $\{X_1 = 1\}$ (respectively, on $\{X_1 = 1\}$) we have $D = 1 + D_{(\pm)}$, where $D_{(+)}$ (respectively, $D_{(-)}$) has under $\mathbb{P}_{s+1}$ (resp., $\mathbb{P}_{s-1}$) the same distribution as D. This leads to

$$\mathfrak{d}(s) \,=\, p\,\big(1 + \mathbb{E}_{s+1}(D)\big) \,+\, q\,\big(1 + \mathbb{E}_{s-1}(D)\big) \,=\, 1 + p\;\mathfrak{d}(s+1) + q\;\mathfrak{d}(s-1)\,,$$

a second-order difference equation that has to be solved subject to the boundary conditions $\mathfrak{d}(0) = \mathfrak{d}(b) = 0$. The solution is

$$\mathfrak{d}(s) \,=\, \frac{s}{q-p} - \frac{b}{q-p}\,\frac{1 - \big(q/p\big)^s}{1 - \big(q/p\big)^b} \quad \text{for } \; p \ne 1/2\,; \qquad \mathfrak{d}(s) \,=\, s(b-s) \quad \text{for } \; p = 1/2\,.$$

Again, if Peter is infinitely rich ($b = \infty$, thus $H_b \equiv \infty$), we get from this expression upon letting $b \to \infty$, formally at least:

$$\mathfrak{d}^*(s) \,:=\, \mathbb{E}(H_0) \,=\, \infty\,, \quad \text{for } \; p \ge 1/2\,;$$
$$= \,\frac{s}{q-p}\,, \quad \text{for } \; p < 1/2\,.$$

Another observation worthy of note emerges: "In a fair game, the poor guy gets wiped out eventually, but this may take a very long time".

Exercise 9.3. For a simple random walk started at the origin, compute the probability of eventual return to the origin for any $p \in (0, 1)$.

9.5 Brownian Motion

To be written...

[39] At the dawn of the subject, back in the middle of the 17th century, Christiaan HUYGENS posed this problem to Pierre DE FERMAT, who solved it. HUYGENS had an awful lot of trouble wrapping his mind around this aspect of FERMAT's solution.

10 Modes of Convergence; Limit Theorems

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and random variables $\{X_n\}_{n\in\mathbb{N}}$, X on it. We shall present in this chapter many different possible notions for the convergence of the sequence $\{X_n\}_{n\in\mathbb{N}}$ to the random variable X, and discuss in a systematic way how these modes of convergence are interrelated.

• **Pointwise Convergence:** This means that the sequence of real numbers $\{X_n(\omega)\}_{n\in\mathbb{N}}$ converges to the real number $X(\omega)$, for every $\omega \in \Omega$. It is the simplest notion of convergence, but also the most stringent.

• **Almost-Everywhere Convergence:** This is the most familiar notion of convergence, meaning that there exists a set $\Omega^* \in \mathcal{F}$ with $\mathbb{P}(\Omega^*) = 1$ and such that the sequence of real numbers $\{X_n(\omega)\}_{n\in\mathbb{N}}$ converges to the real number $X(\omega)$ for every $\omega \in \Omega^*$.

We write then $X_n \longrightarrow X$, a.e. This is the notion of convergence in the strong law of large numbers.

Similarly, we say that the sequence $\{X_n\}_{n\in\mathbb{N}}$ is CAUCHY *a.e.* if the sequence of real numbers $\{X_n(\omega)\}_{n\in\mathbb{N}}$ is CAUCHY for every $\omega \in \Omega^*$: $\lim_{n\to\infty,\, m\to\infty} |X_n(\omega) - X_m(\omega)| = 0$.

• **Convergence in Probability:** The sequence $\{X_n\}_{n\in\mathbb{N}}$ converges to X *in probability*, if for every $\varepsilon > 0$ we have $\mathbb{P}(|X_n - X| > \varepsilon) \to 0$ as $n \to \infty$.

This is the notion of convergence in the weak law of large numbers.

By the same token, we say that the sequence $\{X_n\}_{n\in\mathbb{N}}$ is CAUCHY *in probability,* if for every $\varepsilon > 0$ we have

$$\lim_{m,n\to\infty} \mathbb{P}(|X_n - X_m| > \varepsilon) = 0\,.$$

• **Convergence in $\mathbb{L}^p$:** The sequence $\{X_n\}_{n\in\mathbb{N}} \subset \mathbb{L}^p$ *converges in* $\mathbb{L}^p$ to the random variable $X \in \mathbb{L}^p$ for some $p \in (0,\infty)$, if $\mathbb{E}(|X_n - X|^p) \to 0$ as $n \to \infty$.

We say that $\{X_n\}_{n\in\mathbb{N}}$ is CAUCHY *in* $\mathbb{L}^p$, if we have $\lim_{m,n\to\infty} \mathbb{E}(|X_n - X_m|^p) = 0$.

• **Convergence in Distribution:** The sequence $\{X_n\}_{n\in\mathbb{N}}$ converges to the random variable X *in distribution*, if the sequence of distribution functions $\{F_{X_n}(\cdot)\}_{n\in\mathbb{N}}$ converges to the distribution function $F_X(\cdot)$ at every continuity-point of $F_X(\cdot)$.

This is the notion of convergence in the DE MOIVRE-LAPLACE Theorem and, more generally, in the Central Limit Theorem.

Remark: On Metrizability. Exercise 10.2 below, shows that convergence in probability is equivalent to convergence in the metric

$$\varrho(X,Y) := \mathbb{E}\big(\big|X - Y\big| \wedge 1\big)$$

on the space $\mathbb{L}^0$ of (equivalence classes of) measurable functions. Of course, convergence in $\mathbb{L}^p$, $1 \le p < \infty$ is equivalent to convergence in the metric

$$\varrho(X,Y) := \Big(\mathbb{E}\big(\big|X - Y\big|^p\big)\Big)^{1/p}, \qquad 1 \le p < \infty\,.$$

By contrast, a.e. convergence cannot be metrized. Convergence in distribution can, but we will not go into this important topic in the present course.

For completeness only, we remark that convergence in $\mathbb{L}^p$, $0 < p < 1$ can also be metrized, this time by the metric

$$\varrho(X,Y) \,:=\, \mathbb{E}\big(\big|X-Y\big|^p\big)\,, \qquad 0<p<1\,.$$

10.1 Vague Convergence

We shall say that a sequence of random variables $\{X_n\}_{n\in\mathbb{N}}$ converges to a random variable X *vaguely*, if $\mathbb{E}[\Psi(X_n)] \to \mathbb{E}[\Psi(X)]$ holds as $n \to \infty$ for any bounded, continuous function $\Psi : \mathbb{R} \to \mathbb{R}$.

Similarly, we shall say that a sequence $\{\mu_n\}_{n\in\mathbb{N}}$ of probability measures on $\mathcal{B}(\mathbb{R})$ converges *vaguely* to a probability measure μ on $\mathcal{B}(\mathbb{R})$, if

$$\lim_{n\to\infty} \int_{\mathbb{R}} \Psi\, d\mu_n \,=\, \int_{\mathbb{R}} \Psi\, d\mu \quad \text{holds for every bounded, continuous function } \Psi : \mathbb{R} \to \mathbb{R}\,.$$

Of course, the sequence $\{X_n\}_{n\in\mathbb{N}}$ converges to X vaguely, if and only if the corresponding sequence $\{\mu_n\}_{n\in\mathbb{N}}$ of induced measures $\mu_n = \mathbb{P}\circ X_n^{-1}$ converges vaguely to the induced measure $\mu = \mathbb{P}\circ X^{-1}$.

For instance, if $\{x_n\}_{n\in\mathbb{N}} \subset \mathbb{R}$ and $\lim_{n\to\infty} x_n = x \in \mathbb{R}$, then $\delta_{x_n} =: \mu_n \longrightarrow \mu := \delta_x$ vaguely as $n \to \infty$. Indeed, for every bounded, continuous function $\Psi : \mathbb{R} \to \mathbb{R}$ we have $\int_{\mathbb{R}} \Psi\, d\mu_n = \Psi(x_n) \longrightarrow \Psi(x) = \int_{\mathbb{R}} \Psi\, d\mu$ as $n \to \infty$.

Similarly, if μ_n denotes, for each $n \in \mathbb{N}$, the probability measure induced on the BOREL subsets of the real line by a random variable $Z/\sqrt{n}$ with Gaussian $\mathcal{N}(0, 1/n)$ distribution, then $\mu_n \longrightarrow \delta_0$ vaguely as $n \to \infty$. Indeed, for every bounded, continuous function $\Psi : \mathbb{R} \to \mathbb{R}$ we have

$$\int_{\mathbb{R}} \Psi(x)\, \mathrm{d}\mu_n(x) \,=\, \int_{\mathbb{R}} \Psi(x/\sqrt{n}\,)\, \mathrm{d}\mu_1(x) \,\longrightarrow\, \int_{\mathbb{R}} \Psi(0)\, \mathrm{d}\mu_1(x) \,=\, \Psi(0) \,=\, \int_{\mathbb{R}} \Psi(x)\, \mathrm{d}\delta_0(x)$$

from the dominated convergence theorem.

We note also that, in this last example with $A = \{0\}$, we have $\mu_n(A) = 0$ for every $n \in \mathbb{N}$ but $\delta_0(A) = 1$. Thus, we learn that *vague convergence* $\mu_n \longrightarrow \mu$ *does* not *imply*

$$\lim_{n\to\infty} \mu_n(A) \,=\, \mu(A)\,, \qquad \text{for every } A \in \mathcal{B}(\mathbb{R})\,.$$

This property *does* hold, however, if $\mu(\partial A) = 0$, where $\partial A := \overline{A} \setminus A$ is the "boundary" and $\overline{A}$ the "closure" of (i.e., the smallest closed set that contains) the set A; or if the stronger mode of convergence *in total variation* prevails:

$$\big|\big|\, \mu_n - \mu \,\big|\big| \,:=\, \sup_{B\in\mathcal{B}(\mathbb{R})} \big|\mu_n(B) - \mu(B)\big| \,\longrightarrow\, 0\,, \quad \text{as } n \to \infty\,.$$

We shall see in Theorem 10.3 down the road, that the last two notions of convergence (in distribution, and vague) are equivalent. It is important also to note, that for these two notions the random variables $\{X_n\}_{n\in\mathbb{N}}$, X need *not* be defined on the same probability space. From Exercise 4.23, if $\{X_n\}_{n\in\mathbb{N}}$ converges vaguely to both X and Y, then these two random variables must have the same distribution.

10.2 Relations

Theorem 10.1. Relations Among Different Modes of Convergence. *We have the following implications:*

(i) Convergence $\mathbb{P}-$a.e. $\Rightarrow$ Convergence in Probability $\Rightarrow$ Convergence in Distribution.

(ii) $\sum_{n\in\mathbb{N}} \mathbb{P}(|X_n - X| > \varepsilon) < \infty\,, \forall\ \varepsilon > 0 \implies \mathbb{P}(|X_n - X| > \varepsilon,\ \text{i.o.}) = 0\,,\ \forall\, \varepsilon > 0$

$$\iff \quad X_n \to X\,, \quad \mathbb{P}-\text{a.e.}$$

$$\iff \lim_{k\to\infty} \mathbb{P}\big(\sup_{n\geq k} |X_n - X| > \varepsilon\big) = 0\,, \forall\, \varepsilon > 0.$$

(iii) $\sum_{n\in\mathbb{N}} \mathbb{E}(|X_n - X|^p) < \infty$, for some $p \in (0,\infty) \implies X_n \to X\,, \quad \mathbb{P}-$a.e.

(iv) Convergence in $\mathbb{L}^p$, for some $p \in (0,\infty) \implies$ Convergence in Probability

$$\implies \text{Convergence } \mathbb{P}-\text{a.e. along some subsequence } \{X_{n_k}\}_{k\in\mathbb{N}}.$$

(v) $\mathbb{P}\big(\lim_{n\to\infty} X_n \text{ exists in } \mathbb{R}\big) = 1 \iff \{X_n\}_{n\in\mathbb{N}}$ is CAUCHY a.e.

$$\iff \lim_{n\to\infty} \mathbb{P}\big(\sup_{k\geq 1} |X_{n+k} - X_n| > \varepsilon\big) = 0 \text{ holds for every } \varepsilon > 0\,.$$

Proof: First, we fix $\omega \in \Omega$ and recall the definition of convergence of $\{X_n(\omega)\}_{n\in\mathbb{N}}$ to $X(\omega)$: for all $\varepsilon > 0$, there exists $k \in \mathbb{N}$, such that $|X_n(\omega) - X(\omega)| \leq \varepsilon$ holds for all $n \geq k$. This translates into

$$\begin{aligned} \{X_n \to X\} &:= \Big\{ \omega \in \Omega \,\Big|\, \lim_{n\to\infty} X_n(\omega) = X(\omega) \Big\} \\ &= \bigcap_{\varepsilon>0} \bigcup_{k\in\mathbb{N}} \bigcap_{n\geq k} \{|X_n - X| \leq \varepsilon\} = \bigcap_{\varepsilon>0} C(\varepsilon)\,, \end{aligned} \tag{10.1}$$

where we have set

$$B_k(\varepsilon) := \bigcap_{n\geq k} \{|X_n - X| \leq \varepsilon\} \qquad \text{and} \qquad C(\varepsilon) := \bigcup_{k\in\mathbb{N}} B_k(\varepsilon)\,.$$

Note that

$$\big(C(\varepsilon)\big)^c = \bigcap_{k\in\mathbb{N}} \bigcup_{n\geq k} \big\{|X_n - X| > \varepsilon\big\} \equiv \{|X_n - X| > \varepsilon,\ \text{i.o.}\}\,.$$

Now $X_n \to X$ a.e. $\Leftrightarrow \mathbb{P}(X_n \to X) = 1 \Leftrightarrow \mathbb{P}(C(\varepsilon)) = 1$ for all $\varepsilon > 0$, so we obtain the following important characterization of $\mathbb{P}-$a.e. convergence:

$$\boxed{X_n \to X \ \text{ a.e.} \quad \iff \quad \mathbb{P}(|X_n - X| > \varepsilon\,,\ \text{i.o.}) = 0\,,\ \ \forall\, \varepsilon > 0\,.} \tag{10.2}$$

Since $\mathbb{P}(|X_n - X| > \varepsilon\,,\ \text{i.o.}\,) = 0$ holds if $\sum_{n\in\mathbb{N}} \mathbb{P}(|X_n - X| > \varepsilon) < \infty$ on the strength of the BOREL-CANTELLI lemma, we have also established the first two claims in (ii).

- Returning to (10.1), we note that

$$\mathbb{P}(C(\varepsilon)) = 1 \iff \lim_{k\to\infty} \mathbb{P}(B_k(\varepsilon)) = 1$$

holds for every $\varepsilon > 0$. This gives yet another characterization of $\mathbb{P}-$a.e. convergence:

$$X_n \to X\,,\ \mathbb{P}-\text{a.e.} \iff \forall\,\varepsilon>0\,,\ \lim_{k\to\infty}\mathbb{P}\big(|X_n - X| \le \varepsilon,\ \forall\, n \ge k\big) = 1 \tag{10.3}$$

$$\iff \forall\,\varepsilon>0\,,\ \lim_{k\to\infty}\mathbb{P}\Big(\sup_{n\ge k}|X_n - X| > \varepsilon\Big) = 0\,.$$

It also establishes the last equivalence in (ii), as well as the first implication in (i), in view of the elementary inequality $\mathbb{P}(|X_k - X| > \varepsilon) \le \mathbb{P}(\sup_{n\ge k}|X_n - X| > \varepsilon)$.

• A more direct way of showing that convergence $\mathbb{P}-$a.e. implies convergence in probability, proceeds as follows: the random variables $Y_n = \mathbf{1}_{\{|X_n - X|>\varepsilon\}}$ take values in $[0,1]$ and converge a.e. to zero, so the Dominated Convergence Theorem (applicable because we are working on a finite measure space!) gives $\mathbb{P}(|X_n - X| > \varepsilon) = \mathbb{E}(Y_n) \to 0$ as $n \to \infty$.

• For any $p \in (0,\infty)$, the claim (iii) follows easily from (ii) and the ČEBYŠEV inequality $\sum_{n\in\mathbb{N}}\mathbb{P}(|X_n - X| > \varepsilon) \le \varepsilon^{-p}\sum_{n\in\mathbb{N}}\mathbb{E}(|X_n - X|^p)$. The first part of (iv) is also a simple consequence of ČEBYŠEV's inequality.

• For the second implication in (iv), suppose that the sequence $\{X_n\}_{n\in\mathbb{N}}$ is CAUCHY *in probability*; i.e., for every $\varepsilon > 0$, we have $\mathbb{P}\,(\,|X_n - X_m| > \varepsilon\,) \longrightarrow 0$, as $n \to \infty$, $m \to \infty$.

This is certainly the case, if $\{X_n\}_{n\in\mathbb{N}}$ converges in probability. Then we can select a subsequence $\{Y_k\}_{k\in\mathbb{N}} \equiv \{X_{n_k}\}_{k\in\mathbb{N}} \subseteq \{X_n\}_{n\in\mathbb{N}}$ satisfying

$$\mathbb{P}(E_k) \le \frac{1}{2^k}\,,\ \forall\, k\in\mathbb{N}\,,\quad \text{where } E_k := \big\{|Y_{k+1} - Y_k| > 2^{-k}\big\} = \big\{|X_{n_{k+1}} - X_{n_k}| > 2^{-k}\big\}\ ;$$

to wit, $\mathbb{P}(E_k)$ converges to zero "very fast". Now notice that the event $F_m := \bigcup_{k\ge m} E_k$ satisfies $\mathbb{P}(F_m) \le \sum_{k\ge m}\mathbb{P}(E_k) \le 2^{-m+1}$. In particular, for every $\omega \in \Omega \setminus F_m$:

$$|Y_\ell(\omega) - Y_k(\omega)| \le \sum_{j=\ell}^{k-1}|Y_{j+1}(\omega) - Y_j(\omega)| \le \sum_{j=\ell}^{k-1} 2^{-j} \le 2^{-\ell+1}\,,\quad \forall\, k > \ell > m\,.$$

Thus $\{Y_k(\omega)\}_{k\in\mathbb{N}}$ is a CAUCHY sequence of real numbers, and $Y(\omega) := \lim_{k\to\infty} Y_k(\omega)$ exists in $\mathbb{R}$ for every $\omega \in \bigcup_{m\in\mathbb{N}}(\Omega \setminus F_m) = \Omega \setminus F$, where now the event

$$F := \bigcap_{m\in\mathbb{N}} F_m = \bigcap_{m\in\mathbb{N}}\bigcup_{k\ge m} E_k = \limsup_{k\to\infty} E_k$$

satisfies $\mathbb{P}(F) \le \mathbb{P}(F_m) \le 2^{-m+1}$ for all $m \in \mathbb{N}$, thus $\mathbb{P}(F) = 0$. Defining $Y(\omega) \equiv 0$ on F we see that: $Y_k = X_{n_k} \to Y$, $\mathbb{P}-$a.e.

• Assume now that $X_n \to X$ in probability. Then it is straightforward that

$$\mathbb{P}(X_n \le x) = \mathbb{P}(X_n \le x, |X_n - X| \le \varepsilon) + \mathbb{P}(X_n \le x, |X_n - X| > \varepsilon) \tag{10.4}$$

$$\le\ \mathbb{P}(X \le x + \varepsilon) + \mathbb{P}(|X_n - X| > \varepsilon)$$

holds for every $x \in \mathbb{R}$ and every $\varepsilon > 0$. Similarly, we deduce

$$\mathbb{P}(X \le x - \varepsilon) \le \mathbb{P}(X_n \le x) + \mathbb{P}(|X_n - X| > \varepsilon)\,,$$

and letting $n \to \infty$ we obtain

$$F_X(x-\varepsilon) \le \liminf_{n\to\infty} F_{X_n}(x) \le \limsup_{n\to\infty} F_{X_n}(x+\varepsilon) \le F_X(x+\varepsilon)\,. \tag{10.5}$$

If $F_X(\cdot)$ is continuous at x, we can let $\varepsilon \to 0$ to get $\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$. Thus, convergence in probability implies convergence in distribution; this completes the proof of (i). □

Remark: We note that nothing can be said about $\lim_{n\to\infty} F_{X_n}(x)$ if $F_X(\cdot)$ is discontinuous at x, even if we make the stronger assumption that $X_n \to X$ holds $\mathbb{P}-$a.e. In fact, we have in general

$$\{X < x\} \subseteq \liminf_{n\to\infty}\{X_n < x\} \subseteq \limsup_{n\to\infty}\{X_n < x\} \subseteq \{X \le x\}$$

so difficulties arise whenever $\mathbb{P}(X = x) > 0$. For instance, if we set $\Omega = [0,1)$ with LEBESGUE measure, $X_n = (1/n) \to X = 0$, then 0 is a discontinuity point for $F_X(\cdot)$, and $F_X(0) = 1$ holds while $F_{X_n}(0) = 0$ for all $n \in \mathbb{N}$.

Exercise 10.1. (i) Argue from first principles that convergence in probability implies vague convergence.

(ii) Show by example that convergence in probability does *not* imply a.e. convergence.

(iii) Let us agree to say that a sequence of random variables $\{X_n\}_{n\in\mathbb{N}}$ converges *fast in probability* to another random variable X, if $\sum_{n\in\mathbb{N}} \mathbb{P}(|X_n - X| > \varepsilon) < \infty$ holds for every $\varepsilon > 0$.

In Theorem 10.1 we saw that fast convergence in probability implies a.e. convergence. Show by example that the converse is not true.

Exercise 10.2. Let $\mathbb{L}^0$ denote the space of (equivalence classes of) random variables $X : \Omega \to \mathbb{R}$ on a given probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and define $\varrho(X,Y) := \mathbb{E}\big(|X-Y| \wedge 1\big)$ for any two elements X, Y of $\mathbb{L}^0$.

(i) Show that this defines a metric on the space $\mathbb{L}^0$, and that convergence under this metric is equivalent to convergence in probability.

(ii) Show that, under this metric, $\mathbb{L}^0$ becomes a complete metric space: if a sequence of random variables in CAUCHY in probability, then it converges in probability to some random variable.

Exercise 10.3. Let $X_1, X_2, \cdots$ be independent random variables on a given probability space $(\Omega, \mathcal{F}, \mathbb{P})$, with $X_1 = 0$ and

$$\mathbb{P}\big(X_n = \pm n\big) = \frac{1}{2n\log n}\,, \qquad \mathbb{P}\big(X_n = 0\big) = 1 - \frac{1}{n\log n}\,, \quad n = 2, 3, \cdots\,.$$

Show that the average $\overline{X}_n := (1/n)\sum_{j=1}^n X_j$ converges to zero in $\mathbb{L}^2$, thus also in probability, but *not a.e.*

Exercise 10.4. Let $X_1, X_2, \cdots$ be independent random variables with $\mathbb{P}(X_n = 1) = p_n \in (0,1)$ and $\mathbb{P}(X_n = 0) = 1 - p_n$. Show that $\lim_{n\to\infty} X_n = 0$ holds

- in probability, if and only if $\lim_{n\to\infty} p_n = 0$;
- a.e., if and only if $\sum_{n\in\mathbb{N}} p_n < \infty$.

Exercise 10.5. The Strong Law of Large Numbers for Uncorrelated Random Variables: If $X_1, X_2, \cdots$ are pairwise uncorrelated random variables and $K := \sup_{n\in\mathbb{N}} \mathbb{E}(X_n^2) < \infty$, then

$$\lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} \big(X_j - \mathbb{E}(X_j)\big) \;=\; 0\,, \qquad \text{a.e.}$$

If, furthermore, these random variables $X_1, X_2, \cdots$ have all the same expectation $\mathbb{E}(X_j) = m \in \mathbb{R}$, then

$$\lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} X_j \;=\; m\,, \qquad \text{a.e.}$$

(*Hint:* Use ČEBYŠEV, BOREL-CANTELLI and Theorem 10.1, to establish the result along the subsequence $k_n = n^2$; then argue that "nothing bad happens" between n^2 and $(n+1)^2$.)

Exercise 10.6. The MARKOV-CANTELLI Strong Law of Large Numbers: Consider independent random variables $X_1, X_2, \cdots$ with $\sup_{n\in\mathbb{N}} \mathbb{E}(X_n^4) < \infty$. Show that

$$\lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} \big(X_j - \mathbb{E}(X_j)\big) \;=\; 0\,, \qquad \text{holds a.e.}$$

(*Hint:* Proceed as in the Hint for Exercise 10.5, working now with fourth instead of second powers. Note also that the full strength of independence is not necessary.)

Exercise 10.7. A strengthening of BOREL-CANTELLI: If $A_1, A_2, \cdots$ are pairwise-independent events with $\sum_{n\in\mathbb{N}} \mathbb{P}(A_n) = \infty$, then for

$$Q_n \;:=\; \frac{\sum_{j=1}^{n} \mathbf{1}_{A_j}}{\sum_{j=1}^{n} \mathbb{P}(A_j)} \quad \text{we have} \quad \lim_{n\to\infty} Q_n \;=\; 1 \quad \text{a.e.}$$

(*Hint:* Establish convergence in probability first. Then proceed to show almost-everywhere convergence: first for the subsequence $n_k = \inf\{n \geq 1 \,:\, \sum_{j=1}^{n} \mathbb{P}(A_j) \geq k^2\}$, $k \in \mathbb{N}$, and then for the entire sequence.)

Exercise 10.8. Let $X_1, X_2, \cdots$ be independent random variables with common distribution. Find necessary and sufficient conditions, for the sequence $\big\{\big(X_n/n\big)\big\}_{n\in\mathbb{N}}$ to converge to $X \equiv 0$: (a) a.e.; (b) in probability.

10.3 Ramifications

In Theorem 10.1 there is no mention of whether convergence in distribution can provide any information on any other kind of convergence. This is to be expected, of course, since distinct random variables can have the same distribution!

Perhaps paradoxically, however, convergence in distribution actually implies convergence a.e., if we can *choose both* the representative random variables *and* the probability spaces they are defined on. In fact, we have the following result, which complements Proposition 6.3.

Theorem 10.2. SKOROHOD Representation: *Let F, $\{F_n\}_{n\in\mathbb{N}}$ be probability distribution functions on the real line, and suppose that $\lim_n F_n(x) = F(x)$ holds at every continuity point x of $F(\cdot)$. Then there exists a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and real-valued random variables X, $\{X_n\}_{n\in\mathbb{N}}$ on it, such that $F_X = F$, $F_{X_n} = F_n$ for all $n \in \mathbb{N}$, and $X_n \to X$, $\mathbb{P}-$a.e.*

Proof : Take $\Omega = [0,1]$ with its BOREL sets and LEBESGUE measure $\mathbb{P} \equiv \lambda$, and recall the SKOROHOD construction

$$X_n^+(\omega) = \inf\{x : F_n(x) > \omega\} \geq X_n^-(\omega) = \inf\{x : F_n(x) \geq \omega\}$$

of Proposition 6.3. We have $F_{X_n^\pm} \equiv F_n$, $\mathbb{P}(X_n^+ = X_n^-) = 1$ for every $n \in \mathbb{N}$, and $F_{X^\pm} \equiv F$, $\mathbb{P}(X^+ = X^-) = 1$. Denote by $\mathcal{D}$ the set of discontinuity points of F.

Fix now $\omega \in \Omega$, and take $x \in (X^+(\omega), \infty) \bigcap \mathcal{D}^c$. We have then $x > X^+(\omega)$, so $F(x) > \omega$, and consequently $F_n(x) > \omega$ (hence also $X_n^+(\omega) \leq x$) for all n sufficiently large; thus, it develops that $\limsup_n X_n^+(\omega) \leq x$. Letting $x \downarrow X^+(\omega)$ along $\mathcal{D}^c$ (which is possible, because $\mathcal{D}$ is at most countable) we obtain $\limsup_n X_n^+(\omega) \leq X^+(\omega)$.

In a similar manner, we obtain $\liminf_n X_n^-(\omega) \geq X^-(\omega)$; and this leads to

$$X^+(\omega) \geq \liminf_n X_n^+(\omega) \geq \liminf_n X_n^-(\omega) \geq X^-(\omega).$$

Since $\mathbb{P}(X^+ = X^-) = 1$, the result follows. □

Theorem 10.3. Equivalence of Vague and Distributional Convergence: *Let $\{\mu_n\}$, μ be probability measures on the real line, and $\{F_n\}$, F their corresponding distribution functions.*

Then $\lim_{n\to\infty} F_n(x) = F(x)$ holds at every continuity point x of F, if and only if for all Ψ in the class $\mathcal{C}_b(\mathbb{R})$ of bounded, continuous functions on the real line we have

$$\lim_{n\to\infty} \int_{\mathbb{R}} \Psi \, d\mu_n = \int_{\mathbb{R}} \Psi \, d\mu .$$

In particular, if X and $\{X_n\}_{n\in\mathbb{N}}$ are random variables, then $X_n \to X$ vaguely, if and only if $X_n \to X$ in distribution.

Proof: By the SKOROHOD representation Theorem 10.2, we can view $\{F_n\}$, F as the distribution functions $\{F_{X_n}\}$, F_X of random variables $\{X_n\}$, X such that $X_n(\omega) \to X(\omega)$ for $\mathbb{P}-$a.e. $\omega \in \Omega$; then $\{\mu_n\}$, μ are the corresponding induced measures. Then for $\Psi : \mathbb{R} \to \mathbb{R}$ bounded and continuous, we have

$$\int_{\mathbb{R}} \Psi \, d\mu_n = \mathbb{E}[\Psi(X_n)] \longrightarrow \mathbb{E}[\Psi(X)] = \int_{\mathbb{R}} \Psi \, d\mu$$

by the LEBESGUE Dominated Convergence Theorem. This shows that convergence in distribution implies vague convergence.

To prove the converse, assume that $\lim_{n\to\infty} \int_{\mathbb{R}} \Psi \, d\mu_n = \int_{\mathbb{R}} \Psi \, d\mu$ holds for all bounded, continuous $\Psi : \mathbb{R} \to \mathbb{R}$ (equivalently, that $X_n \to X$ vaguely). Let $x \in \mathbb{R}$, $\delta > 0$, and choose

$h : \mathbb{R} \to [0,1]$ to be continuous, equal to $h(y) = 1$ for $y \le x$, and equal to $h(y) = 0$ for $y > x + \delta$. In particular, $\mathbf{1}_{(-\infty,x]} \le h \le \mathbf{1}_{(-\infty,x+\delta]}$. Evidently,

$$F_n(x) \le \mathbb{E}[\, h(X_n)\,] = \int_{\mathbb{R}} h\, d\mu_n \le F_n(x+\delta), \qquad F(x) \le \mathbb{E}[\, h(X)\,] = \int_{\mathbb{R}} h\, d\mu \le F(x+\delta)\,.$$

Then vague convergence implies

$$\limsup_{n\to\infty} F_n(x) \le \lim_{n\to\infty} \int_{\mathbb{R}} h\, d\mu_n = \int_{\mathbb{R}} h\, d\mu \le F(x+\delta),$$

and hence, using the fact that F is right continuous, $\limsup_{n\to\infty} F_n(x) \le F(x)$.

Now we introduce the function $g(\cdot) = h(\cdot + \delta)$ and observe that we have $\mathbf{1}_{(-\infty,x-\delta]} \le g \le \mathbf{1}_{(-\infty,x]}$, thus

$$F_n(x-\delta) \le \mathbb{E}[\, g(X_n)\,] = \int_{\mathbb{R}} g\, d\mu_n \le F_n(x)\,, \qquad F(x-\delta) \le \mathbb{E}[\, g(X)\,] = \int_{\mathbb{R}} g\, d\mu \le F(x),$$

whence

$$\liminf_{n\to\infty} F_n(x) \ge \lim_{n\to\infty} \int g\, d\mu_n = \int g\, d\mu \ge F(x-\delta), \quad \liminf_{n\to\infty} F_n(x) \ge F(x-).$$

When x is a point of continuity for $F(\cdot)$, we have actually equalities everywhere. □

Remark 10.1. *From Continuous to Smooth Functions:* Suppose we repeat the argument in the above proof, but now with a modified function $\mathfrak{h} : \mathbb{R} \to [0,1]$ defined as follows: $\mathfrak{h}(y) = 1$ for $y < x$, $\mathfrak{h}(y) = 0$ for $y > x + \delta$, and

$$\mathfrak{h}(y) \;=\; \frac{1}{c} \int_y^{x+\delta} e^{\,-1/s(x+\delta-s)}\, ds\,, \qquad c := \int_x^{x+\delta} e^{\,-1/s(x+\delta-s)}\, ds\,.$$

This function has derivatives of all orders which, together with $\mathfrak{h}(\cdot)$, are bounded and continuous. The proof goes through exactly as before. We conclude that the theorem holds even if we replace in its statement the class $\mathcal{C}_b(\mathbb{R})$ by the class $\mathcal{C}_b^\infty(\mathbb{R})$ functions which, along with their derivatives of all orders, are bounded and continuous on the real line.

Exercise 10.9. Vague Convergence is Metrizeable: For two probability measures μ, ν on $\mathcal{B}(\mathbb{R})$, define

$$\varrho(\mu,\nu) := \inf\big\{\varepsilon > 0 : \mu(A) \le \varepsilon + \nu(A^\varepsilon) \text{ and } \nu(A) \le \varepsilon + \mu(A^\varepsilon) \text{ hold for every set } A \in \mathcal{B}(\mathbb{R})\big\}$$

with the notation $A^\varepsilon := \big\{x \in \mathbb{R} : |x - y| < \varepsilon \text{ for some } y \in A\big\}$.

(i) Argue that this defines a distance on the set $\mathcal{P}(\mathbb{R})$ of probability measures on the real line, called "PROKHOROV distance".

(ii) Show that a sequence $\{\mu_n\}_{n\in\mathbb{N}} \subset \mathcal{P}(\mathbb{R})$ converges vaguely to some $\mu \in \mathcal{P}(\mathbb{R})$ if, and only if, $\lim_{n\to\infty} \varrho(\mu,\mu_n) = 0$.

(iii) For any given probability measures μ, ν in $\mathcal{P}(\mathbb{R})$, and $\varepsilon > 0$, we have $\varrho(\mu,\nu) < \varepsilon$ if, and only if, there exist (on some probability space) random variables X, Y with respective distributions μ, ν and $\mathbb{P}(|X - Y| > \varepsilon) < \varepsilon$.

10.4 KOLMOGOROV's Strong Law of Large Numbers

Suppose that we carry out independent copies of the same experiment; or that, in the course of the same experiment, we observe repeatedly independent copies $X_1, X_2, \cdots$ of a certain numerical characteristic X. Then we expect the arithmetic mean or "sample average"

$$\overline{X}_n := \frac{1}{n} \sum_{j=1}^{n} X_j \quad \text{to converge, in some sense, to the ensemble average } \mathbb{E}(X), \tag{10.6}$$

as the number n of observations becomes large ($n \to \infty$). Here by "ensemble average" we mean of course the expectation

$$\mathbb{E}(X) = \int_{\Omega} X(\omega)\, d\mathbb{P}(\omega)\,.$$

For instance, in the special context of coin-tossing with $\Omega = \{0,1\}^{\mathbb{N}}$ and independent $X_j(\omega) = \omega_j$ with common distribution $\mathbb{P}(X_j = 1) = 1 - \mathbb{P}(X_j = 0) = p \in (0,1)$, we expect $\lim_{n\to} \overline{X}_n(\omega) = p$ to hold for a.e. $\omega \in \Omega$. We saw manifestations of this principle in Exercises 10.5 and 10.6.

A celebrated result of KOLMOGOROV (1930) shows that (10.6) holds $\mathbb{P}-$a.e. $\omega \in \Omega$, under the condition $\mathbb{E}(|X|) < \infty$, for independent random variables $X_1, X_2, \cdots$ that have the same distribution as X. A strengthening of this result due to ETEMADI (1981) requires these random variables to be only *pairwise* independent. This stronger version follows right below.

Theorem 10.4. KOLMOGOROV's Strong Law of Large Numbers: *Let $X_1, X_2, \cdots$ be a sequence of pairwise-independent random variables, with the same distribution and $\mathbb{E}(|X_1|) < \infty$. Setting $S_n := \sum_{k=1}^{n} X_k$ and $\overline{X}_n := S_n \,/\, n$, we have*

$$\lim_{n\to\infty} \overline{X}_n(\omega) \,=\, \mathbb{E}(X_1) \quad \textit{for a.e. } \omega \in \Omega\,. \tag{10.7}$$

Let us try to place this result into some context, before working out its proof. Unlike Theorem 7.2, which imposes finite second moments and proves only convergence in probability, this result establishes a.e. convergence while only an integrability (first moment) condition. And unlike earlier results establishing a.s. convergence, such as Exercises 10.5 and 10.6, it does not need to resort to finite fourth, or even second, moments. It does not even need the random variables $X_1,\, X_2,\, \cdots$ to be independent: pairwise independence is sufficient for the result to hold, as shown by ETEMADI (1981).

The original result of KOLMOGOROV (1930) became quickly – that is, within two years of its publication – a special case of a much more general result, the celebrated BIRKHOFF (1932) ergodic theorem. In contrast, Theorem 10.4 is not covered by BIRKHOFF's ergodic theorem.

For a sequence $X_1, X_2, \cdots$ of independent random variables, let us also observe that the KOLMOGOROV Zero-One Law of Theorem 7.6 asserts that the event $\{\lim_{n\to\infty} \overline{X}_n \text{ exists in } \mathbb{R}\}$ has probability either zero or one; Theorem 10.4 proves that this probability is one. Likewise, the KOLMOGOROV Zero-One Law mandates that the random variable $\lim_{n\to\infty} \overline{X}_n$ should then be a.e. equal to a constant; Theorem 10.4 identifies this constant as $m = \mathbb{E}(X_1)$.

Proof of Theorem 10.4 (ETEMADI (1981)): The random variables in each of the sequences $\{X_n^{\pm}\}_{n\in\mathbb{N}}$ of positive and negative parts, are still pairwise independent and identically distributed, so it suffices to prove the theorem with X_n replaced by $X_n^{\pm}$. Thus, we may assume $X_n \geq 0$.

We shall use the *method of truncation,* and set $Z_n = X_n$ when $X_n \leq n$, $Z_n = 0$ otherwise; nore concisely, $Z_n := X_n \, \mathbf{1}_{\{X_n \leq n\}}$. *These random variables are still pairwise independent.* The main step, is then to show that for $\mathbb{P}-$a.e. $\omega \in \Omega$ we have

$$\frac{1}{n}\sum_{k=1}^{n} Z_k(\omega) \;\longrightarrow\; \mathbb{E}(X_1) \qquad \text{as } n \to \infty\,. \tag{10.8}$$

For suppose that (10.8) has been established; we claim that (10.7) follows. Indeed, for $\mathbb{P}-$a.e. $\omega \in \Omega$ we have then

$$X_n(\omega) \;=\; Z_n(\omega)\,, \quad \text{for all } \; n \geq N(\omega) \;\text{ sufficiently large;} \tag{10.9}$$

in other words, "the truncation does no lasting damage to the sequence". This is more than enough to justify the passage from (10.8) to (10.7). Now, in order to argue (10.9), observe

$$\sum_{n\in\mathbb{N}} \mathbb{P}\big(X_n \neq Z_n\big) = \sum_{n\in\mathbb{N}} \mathbb{P}\big(X_n > n\big) = \sum_{n\in\mathbb{N}} \mathbb{P}\big(X_1 > n\big) \leq \mathbb{E}(X_1) < \infty\,,$$

so by BOREL-CANTELLI we have $\mathbb{P}\big(X_n \neq Z_n\,,\ \text{i.o.}\big) = 0$ and (10.9) follows.

• For (10.8), we begin by proving this kind of convergence *first along a suitable subsequence.*e More precisely, let $\alpha > 1$ be arbitrary, and set $k_n = \lfloor \alpha^n \rfloor$, where $\lfloor x \rfloor$ stands for the integer part of the real number x (this recipe defines an increasing sequence for n sufficiently large).

Then, by the pairwise independence of $Z_1, Z_2, \cdots$, the zero-mean random variable

$$U_n \;:=\; \frac{1}{k_n}\sum_{j=1}^{k_n} \big(Z_j - \mathbb{E}(Z_j)\big) \qquad \text{has variance} \qquad \frac{1}{k_n^2}\sum_{j=1}^{k_n} \text{Var}(Z_j) \leq \frac{1}{k_n^2}\sum_{j=1}^{k_n} \mathbb{E}(Z_j^2). \quad {}^{40}$$

Thus, we obtain from the ČEBYŠEV inequality

$$\varepsilon^2 \sum_{n\in\mathbb{N}} \mathbb{P}\left(\big|U_n\big| > \varepsilon\right) \leq \sum_{n\in\mathbb{N}} \frac{1}{k_n^2}\sum_{j=1}^{k_n} \mathbb{E}(Z_j^2) \leq \text{const.} \sum_{j\in\mathbb{N}} \mathbb{E}(Z_j^2) \sum_{\substack{n\in\mathbb{N} \\ \lfloor \alpha^n \rfloor \geq j}} \frac{1}{\lfloor \alpha^n \rfloor^2} \tag{10.10}$$

for arbitrary $\varepsilon > 0$. But this last sum is essentially a convergent geometric series, so its size is that of its largest term j^{-2}. Hence the double sum is dominated by a constant C , times

$$\sum_{j\in\mathbb{N}} \frac{1}{j^2}\,\mathbb{E}(Z_j^2) \;=\; \sum_{j\in\mathbb{N}} \frac{1}{j^2}\,\mathbb{E}\big[\,(X_j)^2\,\mathbf{1}_{\{X_j \leq j\}}\,\big] \;=\; \sum_{j\in\mathbb{N}} \frac{1}{j^2}\,\mathbb{E}\big[\,(X_1)^2\,\mathbf{1}_{\{X_1 \leq j\}}\,\big]\,.$$

[40] This is the only place in the proof, where independence is evoked; and pairwise independence is enough, for arguing that the variance of a sume of random variables is equal to the sum of the individual variances, thus dominated by the sum of the second moments.

This last series can be re-expressed in a "layered" manner, as

$$\sum_{j\in\mathbb{N}} \frac{1}{j^2} \sum_{\ell=0}^{j-1} \mathbb{E}\big[\,(X_1)^2\,\mathbf{1}_{\{\ell\le X_1<\ell+1\}}\,\big] = \sum_{\ell\in\mathbb{N}_0} \mathbb{E}\big[\,(X_1)^2\,\mathbf{1}_{\{\ell\le X_1<\ell+1\}}\,\big] \left(\sum_{j\ge\ell+1} \frac{1}{j^2}\right)$$

$$= C \sum_{\ell\in\mathbb{N}_0} \frac{1}{\ell+1}\, \mathbb{E}\big[\,(X_1)^2\,\mathbf{1}_{\{\ell\le X_1<\ell+1\}}\,\big] \;\le\; C \sum_{\ell\in\mathbb{N}_0} \mathbb{E}\big[\,X_1\cdot\mathbf{1}_{\{\ell\le X_1<\ell+1\}}\,\big] \;=\; C\,\mathbb{E}(X_1) < \infty.$$

Thus, the left-hand side in (10.10) is finite for any $\varepsilon > 0$; in other words, the sequence $\{U_n\}$ converges to zero in probability *fast.*

This, in turn implies, by Theorem 10.1(ii), that we have the a.e. convergence

$$\lim_n \frac{1}{k_n} \sum_{j=1}^{k_n} \big(Z_j - \mathbb{E}(Z_j)\big) \;=\; \lim_n U_n \;=\; 0\,.$$

But

$$\lim_n \frac{1}{k_n} \sum_{j=1}^{k_n} \mathbb{E}(Z_j) \;=\; \lim_n \mathbb{E}(Z_n) \;=\; \lim_n \mathbb{E}\big(X_1\,\mathbf{1}_{\{X_1\le n\}}\big) \;=\; \mathbb{E}(X_1)$$

by Monotone or Dominated Convergence, and then it follows that $\lim_n \big(\sum_{j=1}^{k_n} Z_j\big)/k_n \;=\; \mathbb{E}(X_1)$ holds, a.e.

In other words, we have established the convergence in (10.8) along a suitable subsequence.

• To obtain convergence in (10.8) for the *entire sequence,* we proceed by the so-called *sandwich method.* For every given $k\in\mathbb{N}$, define n_k by $\lfloor\alpha^{n_k}\rfloor \le k < [\alpha^{n_k+1}]$. Since $Z_n \ge 0$, we have

$$\frac{1}{\alpha}\cdot\frac{\sum_{j=1}^{\lfloor\alpha^{n_k}\rfloor} Z_j}{\lfloor\alpha^{n_k}\rfloor} \le \frac{\sum_{j=1}^{\lfloor\alpha^{n_k}\rfloor} Z_j}{\lfloor\alpha^{\,n_k+1}\rfloor} \le \frac{\sum_{j=1}^{k} Z_j}{k} \le \frac{\sum_{j=1}^{\lfloor\alpha^{\,n_k+1}\rfloor} Z_j}{\lfloor\alpha^{n_k}\rfloor} \le \frac{\alpha\cdot\sum_{j=1}^{\lfloor\alpha^{\,n_k+1}\rfloor} Z_j}{\lfloor\alpha^{\,n_k+1}\rfloor}\,.$$

Letting $k\to\infty$ yields

$$\frac{\mathbb{E}(X_1)}{\alpha} \;\le\; \mathrm{liminf}_{k\to\infty}\left(\frac{1}{k}\sum_{j=1}^{k} Z_j\right) \;\le\; \mathrm{limsup}_{n\to\infty}\left(\frac{1}{k}\sum_{j=1}^{k} Z_j\right) \;\le\; \alpha\,\mathbb{E}(X_1)\,;$$

and letting $\alpha\to 1$, we obtain (10.8). □

Theorem 10.5. **GLIVENKO-CANTELLI:** *If $X_1, X_2\cdots$ are independent random variables on $(\Omega,\mathcal{F},\mathbb{P})$ with common distribution μ, then we have the vague convergence of the so-called "empirical distributions" to the true underlying distribution of these random variables, for a.e. $\omega\in\Omega$:*

$$\mu_n(\omega) \;:=\; \frac{1}{n}\sum_{j=1}^{n} \delta_{X_j(\omega)} \;\longrightarrow\; \mu\,, \quad \text{as } n\to\infty\,.$$

Proof: Given any bounded, continuous function $\Psi:\mathbb{R}\to\mathbb{R}$, the strong law of large numbers gives, for a.e. $\omega\in\Omega$:

$$\int_{\mathbb{R}} \Psi\, d\mu_n(\omega) \;=\; \frac{1}{n}\sum_{j=1}^{n} \Psi\big(X_j(\omega)\big) \;\longrightarrow\; \mathbb{E}\,\big[\,\Psi\big(X_1\big)\,\big] \;=\; \int_{\mathbb{R}} \Psi\, d\mu\,, \quad \text{as } n\to\infty\,. \qquad \square$$

Example 10.1. BOREL's Normal Numbers: One of the first strong laws of large numbers was proved by Émile BOREL (1909). He observed that the RADEMACHER functions $\{r_k\}_{k\in\mathbb{N}}$ of Example 7.6 satisfy

$$\lim_{n\to\infty} \frac{r_1(\omega) + \cdots + r_n(\omega)}{n} = 0\,, \quad \text{for } \mathbb{P}-\text{a.e. } \omega \in [0,1)\,, \tag{10.11}$$

where we denote by $\mathbb{P}$ the LEBESGUE measure on the unit interval.

BOREL used essentially the method of proof for the MARKOV and CANTELLI stong laws; cf. Exercises 4.17 and 10.6. Here is the gist of his argument: if $\{f_n\}_{n\in\mathbb{N}}$ is a sequence of non-negative and integrable functions on [0,1), then the convergence of $\sum_{n\in\mathbb{N}} \int_0^1 f_n(\omega)\,d\omega$ implies the convergence of the series $\sum_{\in\mathbb{N}} f_n(\omega)$ for $\lambda-$a.e. $\omega \in [0,1)$. Now take

$$f_n(\omega) = \left(\frac{r_1(\omega) + \cdots + r_n(\omega)}{n}\right)^4$$

and observe that for this choice

$$\int_0^1 f_n(\omega)\,d\omega = \frac{1}{n^4}\left(n + \frac{4!}{2!2!}\frac{n!}{2!(n-2)!}\right) = \frac{1}{n^4}\big(n + 3\,n\,\big(n-1\big)\big)$$

is the general term of a convergent series, thus $\sum_{n\in\mathbb{N}} \int_0^1 f_n(\omega)\,d\omega < \infty$. We conclude that for $\mathbb{P}-$a.e. $\omega \in [0,1)$, we have $\sum_{n\in\mathbb{N}} f_n(\omega) < \infty$ thus also $\lim_{n\to\infty} f_n(\omega) = 0$, and we are done.

In the notation of Example 7.7 we have $r_k = 1 - 2\varepsilon_k$, so (10.11) can be written equivalently

$$\lim_{n\to\infty} \frac{\varepsilon_1(\omega) + \cdots + \varepsilon_n(\omega)}{n} = \frac{1}{2}\,, \quad \text{for } \mathbb{P}-\text{a.e. } \omega \in [0,1)\,. \tag{10.12}$$

In other words, *almost every number* $\omega \in [0,1)$ *has asymptotically the same proportion of 0's and 1's in its binary expansion.* We express this property as *normality to base 2.*

From a probabilistic point of view the statement (10.12) is just the strong law of large numbers applied to the sequence of independent BERNOULLI variables $\varepsilon_1, \varepsilon_2, \cdots$ of Example 7.7, for which $\mathbb{P}(\varepsilon_n = 0) = \mathbb{P}(\varepsilon_n = 1) = 1/2$.

Discussion: Of course, nothing about the particular base 2 is sacrosanct. If $b \geq 2$ is an integer, we also have a unique expansion

$$\omega = \frac{\zeta_1(\omega)}{b} + \frac{\zeta_2(\omega)}{b^2} + \cdots + \frac{\zeta_n(\omega)}{b^n} + \cdots$$

for every $\omega \in [0,1)$, where each digit $\zeta_n(\omega)$ takes values in $\{0, 1, \cdots, b-1\}$. Then one shows that every given digit $k \in \{0, 1, \cdots, b-1\}$ occurs with the same asymptotic frequency in this expansion, namely:

$$\lim_{n\to\infty} \frac{1}{n}\sum_{j=1}^{n} \mathbf{1}_{\{k\}}(\zeta_j(\omega)) = \frac{1}{b}\,, \tag{10.13}$$

for $\mathbb{P}-$a.e. $\omega \in [0,1)$. This is *normality to base b.* But the union of countably many null sets is a null set, so we conclude that (LEBESGUE) *almost every number in* [0,1) *is* **normal**, meaning that it satisfies (10.13) for all digits $k = 0, 1, \cdots, b-1$ and all bases $b \geq 2$.

It is very ironic, that it is actually quite hard to exhibit even one member of this overwhelming majority! No rational number is normal, though it might be normal to a particular basis (for example, $\frac{1}{3} = \frac{0}{2} + \frac{1}{2^2} + \frac{0}{2^3} + \frac{1}{2^4} + \frac{0}{2^5} + \cdots$ is normal to base 2). The simplest known example of a normal number is written in usual decimal notation as

$$0.12345678910111213141516171819202122232425262 7.....$$

by listing all positive integers in succession after the decimal point; but even for this number normality is no trivial matter to establish! □

Example 10.2. Random Growth: Consider independent random variables $X_1, X_2, \cdots$ with values in $(0,\infty)$ and a common distribution with $\varrho := \mathbb{E}(X_1) \in (0,\infty)$.

Then the product $\mathfrak{P}_n := \prod_{i=1}^n X_i$ has expectation $\mathbb{E}(\mathfrak{P}_n) = \varrho^n$: this grows exponentially if $\varrho > 1$, decays exponentially if $\varrho < 1$, and stays flat if $\varrho = 1$.

Does this imply "pathwise" growth $\mathfrak{P}_n(\omega) \approx \varrho^n = e^{n \log \varrho}$ as $n \to \infty$, that is,

$$\lim_{n\to\infty} \frac{1}{n} \log \mathfrak{P}_n(\omega) = \log \varrho\,, \qquad \text{for a.e. } \omega \in \Omega\,?$$

The answer is an emphatic NO! The strong law of large numbers gives

$$\lim_{n\to\infty} \frac{1}{n} \log \mathfrak{P}_n(\omega) = \lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^n \log X_j(\omega) \longrightarrow \mathbb{E}\big(\log X_1\big) =: r \qquad \text{for a.e. } \omega \in \Omega\,,$$

provided, of course, $\log X_1 \in \mathbb{L}^1$. Caution: the JENSEN inequality gives

$$r = \mathbb{E}\big(\log X_1\big) \leq \log\big(\mathbb{E}(X_1)\big) = \log \varrho\,.$$

In other words, the growth rate of the expectation overestimates the "real" (pathwise) growth rate, *unless* there is no randomness.

In fact, for a non-degenerate distribution we might easily get $r < 0 < \log \varrho$, so beware of promises of considerable "expected returns on investment"; they may not translate into positive growth rates.

Example 10.3. Portfolio Choice: Suppose we start out with $\$\,1$ and can invest in two assets: one riskless (money market) with known interest rate $r \geq 0$, the other risky (stock) with random returns $R_1, R_2 \cdots$ from period to period. Suppose also that we follow a "proportional" investment strategy: we invest in every period a fixed, constant proportion $\pi \in [0,1]$ of our wealth in the riskless asset, and the remaining proportion $1-\pi$ in the risky asset. Then the dynamics of our wealth $W_n^{(\pi)}$, $n = 0, 1, \cdots$ are given by

$$W_{n+1}^{(\pi)} = \big(\pi(1+r) + (1-\pi)(1+R_n)\big) \cdot W_n^{(\pi)}\,, \qquad W_0^{(\pi)} = 1\,.$$

How do we select π so as to maximize the long-run growth rate from such (proportional) investment?

In order to be able to say something about this issue, let us assume that the stock returns $R_1, R_2 \cdots$ are independent random variables with common distribution and $\alpha := \mathbb{E}(R_i) > r$ ("risk premium").

Then we are in the situation of Example 10.2 with $X_j^{(\pi)} := \pi(1+r) + (1-\pi)(1+R_j)$ and $W_n^{(\pi)} = \prod_{j=1}^n X_j^{(\pi)}$, thus

$$\mathbb{E}\big(W_n^{(\pi)}\big) \;=\; \big(\pi(1+r) + (1-\pi)(1+\alpha)\big)^n .$$

It is very easy to maximize this expression: just take $\pi = 0$ ("invest everything in the stock, the risky asset").

A more sensible approach would be to look at the long-run rate of return, namely

$$\lim_{n\to\infty} \frac{1}{n} \log\big(W_n^{(\pi)}(\omega)\big) \;=\; \lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^n \log X_j^{(\pi)}(\omega) \;=\; \mathbb{E}\big(\log X_1^{(\pi)}\big) , \qquad \text{for a.e. } \omega \in \Omega$$

from the strong law of large numbers, and try to maximize this over $\pi \in [0,1]$. For instance, with $r = 0$ (a very reasonable assumption these days!) and $\mathbb{P}(R_1 = -1) = \mathbb{P}(R_1 = 1 + 2\delta) = 1/2$ for some $\delta > 0$, so that $\mathbb{E}(R_1) = \delta > 0 = r$.

In this case the long-run rate of return

$$f(\pi) \;:=\; \mathbb{E}\big(\log X_1^{(\pi)}\big) \;=\; \frac{1}{2} \log\Big(\pi\big(1 + 2(1-\pi)(\delta+1)\big)\Big)$$

is maximized at

$$\pi_* \;=\; \frac{1}{2}\Big(1 + \frac{1}{2(\delta+1)}\Big) .$$

This echoes a more general theme (e.g., KARATZAS & KARDARAS (2021)): growth-optimal portfolios are typically mixtures of money market and stock investments ("diversification").

Exercise 10.10. For any bounded, continuous function $f : [0,\infty) \to \mathbb{R}$, compute the limits

$$\lim_{n\to\infty} \int_0^1 \cdots \int_0^1 f\left(\frac{x_1 + \cdots + x_n}{n}\right) dx_1 \cdots dx_n ,$$

$$\lim_{n\to\infty} 2^n \int_0^\infty \cdots \int_0^\infty f\left(\frac{x_1 + \cdots + x_n}{n}\right) e^{-2(x_1+\cdots+x_n)}\, dx_1 \cdots dx_n .$$

10.5 Exchangeability and DE FINETTI's Theorem *

Let us recall now the notation and results of section 7.9, and introduce the following allied notion.

Definition 10.1. Exchangeability: We say that $X_1, X_2, \cdots$ is a sequence of *exchangeable* random variables, if the distribution of $X_{\pi(1)}, X_{\pi(2)}, \cdots$ is the same for every finite permutation π of the natural numbers.

It is clear that $X_1, X_2, \cdots$ are exchangeable, if they are independent and have the same distribution (the "I.I.D. property"); a bit more generally, if they have the I.I.D. property conditional on some non-trivial $\sigma-$algebra. The celebrated result that follows, one of the deepest and most elegant in the entire theory, asserts that the random variables in a given sequence $X_1, X_2, \cdots$ are exchangeable if, and only if, they have the I.I.D. property conditionally on an appropriate non-trivial $\sigma-$algebra, which we describe presently.

Let us denote by $\mathcal{P}(\mathbb{R})$ the collection of probability measures on $\mathcal{B}(\mathbb{R})$ and equip it with the topology of vague convergence from subsection 10.1: a sequence $\{\nu_n\}_{n\in\mathbb{N}}$ in $\mathcal{P}(\mathbb{R})$ converges *vaguely* to ν in $\mathcal{P}(\mathbb{R})$, if

$$\lim_{n\to\infty} \int_{\mathbb{R}} \Psi \, \mathrm{d}\nu_n \;=\; \int_{\mathbb{R}} \Psi \, \mathrm{d}\nu \quad \text{holds for every bounded, continuous function } \; \Psi : \mathbb{R} \to \mathbb{R} \, .$$

This topology renders $\mathcal{P}(\mathbb{R})$ a complete, separable metric space (e.g., PARTHASARATHY (1967) or BILLINGSLEY (1968)), so regular conditional probabilities exist in this context on account of Theorem 8.1. We shall use them below without comment.

For a sequence of random variables $X_1, X_2, \cdots$ we denote now by $\boldsymbol{\mu}_n = (1/n) \sum_{j=1}^n \boldsymbol{\delta}_{X_j}$ the *empirical measure* of $X_1, \cdots, X_n$; that is, the mapping $\boldsymbol{\mu}_n : \Omega \to \mathcal{P}(\mathbb{R})$ which satisfies, for every bounded, continuous function $\Psi : \mathbb{R} \to \mathbb{R}$, the property

$$\int_{\mathbb{R}} \Psi \, \mathrm{d}\boldsymbol{\mu}_n(\omega) \;=\; \frac{1}{n} \sum_{j=1}^n \Psi(X_j(\omega)) \, , \quad \forall \, \omega \in \Omega \, .$$

Theorem 10.6. DE FINETTI (1937): *Suppose that $X_1, X_2 \cdots$ are exchangeable random variables, and consider the measurable mapping $\boldsymbol{\mu} : \Omega \to \mathcal{P}(\mathbb{R})$ which provides the conditional distribution of X_1, given the permutation-invariant $\sigma-$algebra $\mathcal{E}$ from section 7.9. Then:*

(i) The sequence $\big(\boldsymbol{\mu}_n(\omega)\big)_{n\in\mathbb{N}}$ of empirical measures converges vaguely to the measure $\boldsymbol{\mu}(\omega)$, for $\mathbb{P}-$a.e. $\omega \in \Omega$.

(ii) For every $k \in \mathbb{N}$ and bounded, continuous functions $\Psi_1, \cdots, \Psi_k$, and denoting by $\mathcal{M} = \sigma(\boldsymbol{\mu})$ the $\sigma-$algebra generated by the measurable mapping $\boldsymbol{\mu}$ in (i), we have

$$\mathbb{E}\Big(\prod_{j=1}^k \Psi_j(X_j) \,\Big|\, \mathcal{M} \Big)(\omega) \;=\; \prod_{j=1}^k \int_{\mathbb{R}} \Psi_j \, \mathrm{d}\boldsymbol{\mu}(\omega) \, , \qquad \mathbb{P} - \textit{a.e.} \;\; \omega \in \Omega \, .$$

In words: Conditionally on $\mathcal{M}$, the random variables $X_1, X_2, \cdots$ are independent with common distribution $\boldsymbol{\mu}$.

Proof: (Taken from KARDARAS (2021).) We recall from section 7.9 the $\sigma-$algebra $\mathcal{E}_n$ of events invariant under permutations of the first n coordinates, and the "overall" invariant $\sigma-$algebra $\mathcal{E} = \bigcap_{n\in\mathbb{N}} \mathcal{E}_n$. Since $\boldsymbol{\mu}$ is the conditional distribution of X_1 given $\mathcal{E}$, we have $\mathbb{E}[h(X_1)\,|\,\mathcal{E}] = \int_{\mathbb{R}} h\,\mathrm{d}\boldsymbol{\mu}$ for every bounded, BOREL measurable $h : \mathbb{R} \to \mathbb{R}$.

By exchangeability, $\boldsymbol{\mu}$ is also the conditional distribution of X_j given the invariant $\sigma-$algebra $\mathcal{E}$, for every $j \in \mathbb{N}$. From the Lemma that follows, we have $\mathbb{E}[h(X_1)\,|\,\mathcal{E}_n] = \mathbb{E}[h(X_j)\,|\,\mathcal{E}_n]$ for every $1 \leq j \leq n$ and thus

$$\mathbb{E}\,[h(X_1)\,|\,\mathcal{E}_n] = \frac{1}{n}\sum_{j=1}^{n} \mathbb{E}\,[h(X_j)\,|\,\mathcal{E}_n] = \mathbb{E}\left[\int_{\mathbb{R}} h\,\mathrm{d}\boldsymbol{\mu}_n \,\Big|\, \mathcal{E}_n\right] = \int_{\mathbb{R}} h\,\mathrm{d}\boldsymbol{\mu}_n\,,$$

because $\boldsymbol{\mu}_n$ is $\mathcal{E}_n-$measurable. Proposition 8.5 now gives the *Strong Law of Large Numbers* (or conditional GLIVENKO-CANTELLI Theorem) *for sequences of exchangeable random variables* in the form of the $\mathbb{P}-$a.e. equality

$$\lim_{n\to\infty} \frac{1}{n}\sum_{j=1}^{n} h(X_j) \;=\; \lim_{n\to\infty} \int_{\mathbb{R}} h\,\mathrm{d}\boldsymbol{\mu}_n \;=\; \lim_{n\to\infty} \mathbb{E}[h(X_1)\,|\,\mathcal{E}_n] \;=\; \mathbb{E}[h(X_1)\,|\,\mathcal{E}\,] = \int_{\mathbb{R}} h\,\mathrm{d}\boldsymbol{\mu}\,.$$

This holds for every such h in a countable generating collection of bounded, continuous functions, so the first claim is established.

We move on now to the second claim, and try to establish it by induction on $k \in \mathbb{N}$. With $k = 1$, this follows from part (i) and from the fact that $\boldsymbol{\mu}$ is $\mathcal{E}-$measurable. Suppose that the claim holds for some $k \in \mathbb{N}$. For bounded, continuous $\Psi_1, \cdots, \Psi_k, \Psi_{k+1}$ we write

$$x^k = (x_1, \cdots, x_k)\,, \qquad \Phi^k(x^k) = \Psi_1(x_1)\cdots\Psi_k(x_k)$$

to simplify typography. The induction hypothesis amounts to

$$\mathbb{E}\big[\Phi^k\big(X^k\big)\,\big|\,\mathcal{M}\big] = \prod_{j=1}^{k} \int_{\mathbb{R}} \Psi_j\,\mathrm{d}\boldsymbol{\mu}\,,$$

and we note that $\mathbb{E}[\Phi^k(X^k)\,\Psi_{k+1}(X_{k+1})\,|\,\mathcal{E}_n] = \mathbb{E}[\Phi^k(X^k)\,\Psi_{k+1}(X_i)\,|\,\mathcal{E}_n]$ holds for $k < i \leq n$ on account of Lemma 10.1 below. Therefore,

$$\begin{aligned}
\mathbb{E}\,[\Phi^k(X^k)\,\Psi_{k+1}(X_{k+1})\,|\,\mathcal{E}_n] &= \frac{1}{n-k}\,\mathbb{E}\left[\Phi^k(X^k)\sum_{j=k+1}^{n} \Psi_{k+1}(X_j)\,\Big|\,\mathcal{E}_n\right] \\
&= \frac{n}{n-k}\,\mathbb{E}\left[\Phi^k(X^k)\int_{\mathbb{R}} \Psi_{k+1}\,\mathrm{d}\boldsymbol{\mu}_n\,\Big|\,\mathcal{E}_n\right] - \frac{k}{n-k}\,\mathbb{E}\left[\Phi^k(X^k)\int_{\mathbb{R}} \Psi_{k+1}\,\mathrm{d}\boldsymbol{\mu}_k\,\Big|\,\mathcal{E}_n\right] \\
&= \frac{n}{n-k}\,\mathbb{E}\left[\Phi^k(X^k)\,\big|\,\mathcal{E}_n\right]\int_{\mathbb{R}} \Psi_{k+1}\,\mathrm{d}\boldsymbol{\mu}_n - \frac{k}{n-k}\,\mathbb{E}\left[\Phi^k(X^k)\int_{\mathbb{R}} \Psi_{k+1}\,\mathrm{d}\boldsymbol{\mu}_k\,\Big|\,\mathcal{E}_n\right],
\end{aligned}$$

again because $\boldsymbol{\mu}_n$ is $\mathcal{E}_n-$measurable. The random variable inside the last conditional expectation is bounded, so the last term goes to zero as $n \to \infty$. Passing to this limit and invoking Proposition 8.5, we obtain

$$\mathbb{E}\left[\Phi^k(X^k)\,\Psi_{k+1}(X_{k+1})\,\middle|\,\mathcal{E}\right] \;=\; \mathbb{E}\left[\Phi^k(X^k)\,\middle|\,\mathcal{E}\right]\,\int_{\mathbb{R}} \Psi_{k+1}\,\mathrm{d}\boldsymbol{\mu}\,;$$

whereas, conditioning further with respect to $\mathcal{M} = \sigma(\boldsymbol{\mu})$ leads to

$$\mathbb{E}\left[\Phi^k(X^k)\,\Psi_{k+1}(X_{k+1})\,|\,\mathcal{M}\right] \;=\; \mathbb{E}\left[\Phi^k(X^k)\,\middle|\,\mathcal{M}\right]\,\int_{\mathbb{R}} \Psi_{k+1}\,\mathrm{d}\boldsymbol{\mu} \;=\; \prod_{j=1}^{k}\int_{\mathbb{R}} \Psi_j\,\mathrm{d}\boldsymbol{\mu}\cdot\int_{\mathbb{R}} \Psi_{k+1}\,\mathrm{d}\boldsymbol{\mu}$$

from the induction hypothesis. This completes the argument. □

Lemma 10.1. *For any* $n \in \mathbb{N}$, *bounded* BOREL*-measurable* $\Psi : \mathbb{R}^n \to \mathbb{R}$, *and bijection* $\pi : \{1,\cdots,n\} \to \{1,\cdots,n\}$, *we have*

$$\mathbb{E}\big[\Psi\big(X_{\pi(1)},\cdots,X_{\pi(n)}\big)\,\big|\,\mathcal{E}_n\big] \;=\; \mathbb{E}\big[\Psi\big(X_1,\cdots,X_n\big)\,\big|\,\mathcal{E}_n\big].$$

Proof: We extend π to a bijection on $\mathbb{N}$, setting $\pi(j) = j$ for $j > n$. For any bounded BOREL-measurable function $\gamma : \mathbb{R}^{\mathbb{N}} \to \mathbb{R}$ which is invariant under permutations of the first n coordinates, exchangeability gives

$$\begin{aligned}\mathbb{E}\big[\Psi\big(X_1,\cdots,X_n\big)\,\gamma(X_1,X_2,\cdots)\big] \;&=\; \mathbb{E}\big[\Psi\big(X_{\pi(1)},\cdots,X_{\pi(n)}\big)\,\gamma(X_{\pi(1)},X_{\pi(2)},\cdots)\big]\\ &=\; \mathbb{E}\big[\Psi\big(X_{\pi(1)},\cdots,X_{\pi(n)}\big)\,\gamma(X_1,X_2,\cdots)\big]\end{aligned}$$

and the claim follows. □

Example 10.4. Exchangeable Coin Tosses: Here is the simplest setting for Theorem 10.6. Suppose $X_1, X_2, \cdots$ are exchangeable random variables, with values in $\{0,1\}$. Then there is a probability distribution function F supported on the unit interval [0,1], such that for every $n \in \mathbb{N}$, $(x_1,\cdots,x_n) \in \{0,1\}^n$, and with $s = \sum_{j=1}^n x_j$, we have

$$\mathbb{P}\big(X_1 = x_1,\cdots,X_n = x_n\big) \;=\; \int_0^1 p^s\big(1-p\big)^{n-s}\,\mathrm{d}F(p)\,.$$

In other words: Even if you find unpalatable the assumption that different tosses of a coin are independent and have the same probability p of "success" (coming up heads), but you are willing to accept that they are exchangeable, you have to accept that their distribution is a mixture of independent BERNOULLI variables with some "latent" distribution F on the probability of success.

Exercise 10.11. If the random variables $X_1, X_2, \cdots$ are exchangeable and square-integrable, argue that $\mathbb{E}(X_1X_2) \geq 0$.

Exercise 10.12. In the context of Lemma 10.1 and with $\mathcal{M} = \sigma(\boldsymbol{\mu})$, $\mathcal{M}_n = \sigma(\boldsymbol{\mu}_n)$, show that

$$\mathbb{E}\big[\Psi\big(X_1,\cdots,X_n\big)\,\big|\,\mathcal{M}_n \vee \mathcal{M}\big] \;=\; \mathbb{E}\big[\Psi\big(X_1,\cdots,X_n\big)\,\big|\,\mathcal{M}_n\big]$$

holds for any $n \in \mathbb{N}$. In statistical terminology: the empirical measure $\boldsymbol{\mu}_n$ is a *sufficient statistic* for $\boldsymbol{\mu}$.

Exercise 10.13. For a sequence of exchangeable random variables $X_1, X_2, \cdots$, the $\sigma-$algebra $\mathcal{M} = \sigma(\boldsymbol{\mu})$, the $\sigma-$algebra $\mathcal{E}$ of exchangeable events, and the tail $\sigma-$algebra $\mathcal{T}$, are the same.

11 The Central Limit Theorem

The DE MOIVRE-LAPLACE limit theorem of Exercise 3.8 can be generalized very broadly, to yield one of the most important, indeed "central", results in the Theory of Probability. We shall state and discuss this generalization in the present section, but defer its proof to later chapters.

Theorem 11.1. Central Limit Theorem: *Let $X_1, X_2, \cdots$ be independent random variables random variables with common distribution and $\mathbb{E}(|X_1|^2) < \infty$.*

If $m := \mathbb{E}(X_1)$, $\sigma := \sqrt{Var\,(X_1)} > 0$ are the expectation and standard deviation of the distribution , and $S_n = \sum_{k=1}^n X_k$, then in the notation of (3.12) we have as $n \to \infty$:

$$\mathbb{P}\left[\, a \,\leq\, \frac{S_n - nm}{\sigma\,\sqrt{n}} \,\leq\, b \,\right] \;\longrightarrow\; \Phi(b) - \Phi(a) \;=\; \int_a^b \frac{e^{-x^2/2}}{\sqrt{2\pi}}\,\mathrm{d}x \tag{11.1}$$

for $-\infty \leq a < b \leq \infty$. In other words, the random sequence $\frac{1}{\sigma\sqrt{n}} \sum_{k=1}^n \big(X_k - m\big)\,,\ n \in \mathbb{N}$ converges vaguely to a standard normal random variable.

• **Heuristic Discussion: Small Deviations in the Law of Large Numbers**. From the Strong Law of Large Numbers we know that, for n large, we have $\overline{X}_n := (S_n \,/\, n\,) \;\sim\; m = \mathbb{E}(X_1)$, and we are led to ask the question:

What is the order of "small fluctuations" (deviations) of the "sample average" $\overline{X}_n$ around the "ensemble average" m ?

In other words, can we find a positive sequence $f(n) \downarrow 0$ (as $n \to \infty$) such that, for all n enough, we have $|\overline{X}_n - m| = O(1) \cdot f(n)$ almost everywhere, or equivalently

$$\mathbb{P}\Big(|\overline{X}_n - m| \leq K\, f(n)\Big) \;=\; 1, \quad \text{for some } K \in (0,\infty)\,?$$

Requiring that this probability be equal to one, turns out to be too much to ask for; nevertheless, according to the Central Limit Theorem

$$\mathbb{P}\left(\frac{\overline{X}_n - m}{f(n)} \,\in\, A\right) \;\sim\; \int_A \frac{e^{-x^2/2}}{\sqrt{2\pi}}\,\mathrm{d}x\,, \quad A \in \mathcal{B}(\mathbb{R})\,, \tag{11.2}$$

we can manage to *make this probability very high,* provided that we choose the order of "small deviations" to be

$$f(n) \;=\; \sigma \,/\, \sqrt{n}\,.$$

Indeed, with $A = [-K, K]$, the probability of (11.2) is approximately $2\,\Phi(K) - 1$ for n large; for instance, this quantity equals 98.76% for $K = 2.5$, and 99.74% for $K = 3$. These numbers are readily available from the ubiquitous tables of the standard normal distribution.

It is rather remarkable that in all of this, only two characteristics from the underlying common distribution of $X_1, X_2, \cdots$ matter at all: the expectation $m = \mathbb{E}(X_1)$ and the variance $\sigma^2 = \mathrm{Var}(X_1)$. This "universality" feature is one of the things that make the Central Limit Theorem such an important tool in applications. It took almost 200 years for this deep fact to be fully realized.

Illustration: *You are in charge of booking policy for* ICAROS *Airlines, which flies jumbo jets with* $K = 555$ *passenger seats. Industry-wide research suggests that ticketed customers actually show up with probability* $p = 0.9$. *How much leeway do you have in overbooking the plane, while still running a risk of no more than* $\alpha = 0.05$ *of having to face, and compensate, irate customers?*

Let us suppose you issue and sell N tickets ($N > K$) and assume that the actions of individual customers are independent; this can be debated, of course, but we have to start somewhere. Then the number S_N of customers who actually do show up, has $\mathrm{Bin}(K, p)$ distribution. We need to select N so that $\mathbb{P}(S_N > K) \leq \alpha$. The actual binomial computation is very arduous, so we resort to Theorem 11.1, which gives

$$\mathbb{P}(S_N > K) = \mathbb{P}\left(\frac{S_N - Np}{\sqrt{Np(1-p)}} > \frac{K - Np}{\sqrt{Np(1-p)}} \right) \approx 1 - \Phi\left(\frac{K - Np}{\sqrt{Np(1-p)}} \right) \leq \alpha\,.$$

With the given values of the parameters, and with help from tables of the Gaussian distribution, this gives $N \leq 602$. □

The rate of convergence in the relation (11.1) of the Central Limit Theorem is quite slow, namely of the order $(1/\sqrt{n}\,)$. This can be seen most clearly by considering the *symmetric* BERNOULLI *distribution* $\mathbb{P}(X_1 = \pm 1) = 1/2$ in Theorem 11.1; in this case it is straightforward to see, using the STIRLING formula of Exercise 3.7, that

$$\big|\mathbb{P}(S_{2n} < 0) - \Phi(0)\big| \;=\; \frac{1}{2}\,\mathbb{P}(S_{2n} = 0) \;=\; \frac{(2n)!}{2\,(n!)^2}\left(\frac{1}{2}\right)^{2n} \;\sim\; \frac{1}{\sqrt{2\pi\,(2n)}}\,.$$

The following result shows that this order of magnitude is typical.

Theorem 11.2. **BERRY-ESSEEN:** *With the same assumptions and notation as in Theorem 11.1 and with the additional condition* $\mathbb{E}(|X_1|^3) < \infty$, *we have*

$$\sup_{x \in \mathbb{R}} \left| \mathbb{P}\left[\frac{S_n - nm}{\sigma\,\sqrt{n}} \;\leq\; x \right] \;-\; \Phi(x) \right| \;\leq\; \frac{C \cdot \mathbb{E}(|X_1|^3)}{\sigma^3\,\sqrt{n}}\,, \quad \forall\, n \in \mathbb{N}$$

for some $C \in (0, \infty)$ *which is universal, that is, does not depend on the distribution of* X_1.

The example preceding Theorem 11.2 makes clear that $C \geq (1\,/\,\sqrt{2\pi}\,)$, and it can be shown that $C < 0.8$. For a proof of this result, see for instance BOLTHAUSEN (1984). □

11.1 LINDEBERG-FELLER Theory

Let us recall from Example 7.8 the number $W_n = \sum_{k=1}^n \mathbf{1}_{A_k}$ of records set by day $t = n$ in a sequence of independent observation drawn from the same, continuous distribution function. We found there that the mean and the variance of this random variable grow at the same rate

$$\mathbb{E}(W_n) \,=\, \sum_{k=1}^{n}(1/k) \,\sim\, \log n\,, \qquad \text{Var}(W_n) \,=\, \sum_{k=1}^{n}(k-1)/k^2 \,\sim\, \log n$$

as $n \to \infty$. This *sublinear* growth results form the fact that W_n is a sum of independent, but not identically distributed, random variables. Is it still the case that

$$\frac{W_n - \log n}{\sqrt{\log n}} \quad \text{converges in distribution to the standard Gaussian ?}$$

Clearly, Theorem 11.1 is inadequate for answering this question. We need a theory that is able to handle *independent* but *non-identically-distributed* random variables. Such a theory was provided by LINDEBERG (1922) and FELLER (1935).

This theory is of fundamental importance in understanding the far-reaching applicability of what we might call with WALSH (2012) the "central limit principle". Nature supplies us with endless varieties of random quantities, but does not always oblige to give them identical distributions (just as right above). In practice, we have only a very vague idea of these distributions. The gist of the fundamental result that follows, Theorem 11.3 below, is that **the sum of a large number of small, independent random variables, is nearly normally distributed.** The word "small" here is meant to indicate that no individual term can dominate the others, in the sense of being nearly equal to their sum.

Let us begin with some definitions and observations, that will give mathematical substance to these remarks.

We shall say that a double array $\{Y_{nj}\}_{j=i,\cdots,k_n;\, n\in\mathbb{N}}$ is *Uniformly Asymptotically Negligible* (UAN), if

$$\lim_{n\to\infty} \max_{1\le j\le k_n} \mathbb{P}\big(|Y_{nj}| > \varepsilon\big) \,=\, 0 \tag{11.3}$$

holds for every $\varepsilon > 0$. It is rather straightforward to see that, if the variables in the double array are square-integrable, then

$$\lim_{n\to\infty} \max_{1\le j\le k_n} \text{Var}(Y_{nj}) \,=\, 0 \tag{11.4}$$

is a sufficient condition for (11.3).

Similarly, *if the variables in the array have the same distribution μ, then the triangular array $\{\widetilde{Y}_{nj} = Y_{nj}/n\}_{j=i,\cdots,k_n,\, n\in\mathbb{N}}$ is UAN.* Indeed, we have then

$$\mathbb{P}(|\widetilde{Y}_{nj}| > \varepsilon) \,=\, \mathbb{P}(|Y_{nj}| > \varepsilon n) \,=\, \mu(A_n) \longrightarrow 0\,, \quad \text{where } A_n := \{x\in\mathbb{R} \,:\, |x| > \varepsilon n\}\,.$$

(For the 'necessity' of considering double arrays in Central Limit Theory, we refer to the Example 11.2 below.)

Lemma 11.1. Conditions of LINDEBERG and FELLER: *Consider a sequence* $\{X_n\}_{n\in\mathbb{N}}$ *of square-integrable random variables, denote by* μ_n *the distribution of* X_n*, and set*

$$m_n := \mathbb{E}(X_n)\,, \quad \sigma_n^2 := \mathit{Var}\,(X_n)\,, \quad s_n^2 := \sigma_1^2 + \cdots + \sigma_n^2\,.$$

We shall say that the sequence satisfies the LINDEBERG condition *if, as* $n \to \infty$*,*

$$L_n(\varepsilon) \;:=\; \frac{1}{s_n^2} \sum_{j=1}^{n} \int_{\{x\in\mathbb{R}\,:\,|x-m_j|>\varepsilon s_n\}} \big(x-m_j\big)^2\,\mathrm{d}\mu_j(x) \;\longrightarrow\; 0 \tag{11.5}$$

holds for every $\varepsilon > 0$*; and that it satisfies the* FELLER condition*, if*

$$\lim_{n\to\infty}\; \frac{1}{s_n^2}\,\left(\max_{1\le j\le n} \sigma_j^2\right) \;=\; 0\,. \tag{11.6}$$

Then condition (11.5) implies condition (11.6), and the latter implies that

$$\textit{the double array} \quad Y_{nj} \;:=\; \frac{1}{s_n}\big(X_j - m_j\big)\,,\; 1\le j\le n\,,\; n\in\mathbb{N} \quad \textit{is UAN.} \tag{11.7}$$

Indeed, let us observe that, for every $\varepsilon > 0$ and $1 \le j \le n$, the variance

$$\sigma_j^2 \;=\; \int_{\{x\in\mathbb{R}\,:\,|x-m_j|\le\varepsilon s_n\}} \big(x-m_j\big)^2\,\mathrm{d}\mu_j(x) \;+\; \int_{\{x\in\mathbb{R}\,:\,|x-m_j|>\varepsilon s_n\}} \big(x-m_j\big)^2\,\mathrm{d}\mu_j(x)$$

is dominated by

$$(\varepsilon s_n)^2 + \int_{\{x\in\mathbb{R}\,:\,|x-m_j|>\varepsilon s_n\}} \big(x-m_j\big)^2\,\mathrm{d}\mu_j(x) \le (\varepsilon s_n)^2 + s_n^2\,L_n(\varepsilon)\,;$$

thus, the expression $s_n^{-2}\,\big(\max_{1\le j\le n}\sigma_j^2\big)$ in (11.6) is in turn dominated by $\varepsilon^2 + L_n(\varepsilon)$ and the implication (11.5) $\Rightarrow$ (11.6) follows.

On the other hand, the FELLER condition implies (11.4) for the sequence in (11.7), and the *UAN* property claimed there follows easily.

Lemma 11.2. *The* LINDEBERG *condition is satisfied, if the random variables* $\{X_n\}_{n\in\mathbb{N}}$
(i) are I.I.D.; or if they
(ii) are uniformly bounded (that is, $|X_n(\omega)| \le M\,,\ \forall\,(n,\omega)$ *holds for some real number* M*) and satisfy* $\lim_{n\to\infty} s_n^2 = \infty$*; or if they*
(iii) satisfy, for some $\delta > 0$*, the* LYAPUNOV *Condition*

$$\lim_{n\to\infty} \sum_{j=1}^{n} \mathbb{E}\,\big|\,Y_{nj}\,\big|^{2+\delta} \;=\; \lim_{n\to\infty}\,\left(\frac{1}{s_n^{2+\delta}} \sum_{j=1}^{n} \mathbb{E}\,\big|\,X_j - m_j\,\big|^{2+\delta}\right) \;=\; 0\,. \tag{11.8}$$

Indeed, in the case of independent random variables with a common distribution,

$$L_n(\varepsilon) = \frac{n}{n\sigma^2} \int_{\{x\in\mathbb{R}\,:\,|x-m|>\varepsilon\sigma\sqrt{n}\,\}} \big(x-m\big)^2 \,\mathrm{d}\mu(x) \;=\; \frac{1}{\sigma^2}\,\mathbb{E}\,\big[\,\big(X_1-m\big)^2\,\mathbf{1}_{\{|X_1-m|>\varepsilon\sigma\sqrt{n}\,\}}\,\big]$$

converges to zero as $n \to \infty$, by dominated convergence. In case (ii), the quantity $L_n(\varepsilon)$ is dominated by

$$\frac{4M^2}{s_n^2}\,\sum_{j=1}^{n}\mathbb{P}\big(\,|\,X_j-m_j\,|>\varepsilon s_n\big) \;\le\; \frac{4M^2}{s_n^2}\,\sum_{j=1}^{n}\frac{\sigma_j^2}{(\varepsilon s_n)^2} \;=\; \frac{4M^2}{\varepsilon^2 s_n^2} \;\longrightarrow\; 0 \quad \text{as } n\to\infty\,.$$

Finally, under the LYAPUNOV condition, the estimate

$$L_n(\varepsilon) \;\le\; \frac{1}{\varepsilon^\delta s_n^{2+\delta}}\,\sum_{j=1}^{n}\int_{\{x\in\mathbb{R}\,:\,|x-m_j|>\varepsilon s_n\}}\big|x-m_j\big|^{2+\delta}\,\mathrm{d}\mu_j(x) \;\le\; \frac{\sum_{j=1}^{n}\,\mathbb{E}\,\big|\,X_j-m_j\,\big|^{2+\delta}}{\varepsilon^\delta s_n^{2+\delta}}$$

shows that condition (11.5) is satisfied.

Theorem 11.3. LINDEBERG-FELLER CLT for Random Variables: *For a sequence $\{X_n\}_{n\in\mathbb{N}}$ of independent, square-integrable random variables with positive variances, let us introduce as in (11.7) the double array*

$$Y_{nj} \;:=\; \frac{1}{s_n}\big(X_j-m_j\big)\,,\qquad 1\le j\le n\,,\quad n\in\mathbb{N}\,.$$

Then the following are equivalent:

(i) The Central Limit Theorem (CLT) holds, namely, the sequence

$$\sum_{j=1}^{n}Y_{nj} \;=\; \frac{1}{s_n}\,\sum_{j=1}^{n}\big(X_j-m_j\big) \quad \textit{converges in distribution to the standard Gaussian} \tag{11.9}$$

as $n\to\infty$; and so does the FELLER *condition (11.6).*

(ii) The Central Limit Theorem (CLT) holds; and so does the UAN condition (11.7).

(iii) The LINDEBERG *condition (11.5) holds.*

This result is very significant: it points out that, for the validity of the Central Limit Theorem *(CLT)*, all one really needs is independence of the variables $\{X_n\}_{n\in\mathbb{N}}$, along with a condition (to wit, (11.7)) guaranteeing that *no single one from among the random variables $X_1,\cdots,X_n$ is big enough to dominate the sum in (11.9).* The variables need *not* have the same distribution.

Example 11.1. Let us continue the discussion of Example 7.8 that we started at the beginning of this subsection, and the number

$$W_n \;=\; \sum_{k=1}^{n}\mathbf{1}_{A_k}$$

of records set by day $t = n$ in a sequence of independent observations $\xi_1, \xi_2, \cdots$ drawn from the same, continuous distribution function.

It was shown in our discussion of Example 7.8, that the events $A_k := \{\xi_k > \max_{1\le j\le k-1} \xi_j\}$ are independent, but *not* identically distributed: $\mathbb{P}(A_k) = 1/k = m_k$ for all $k \in \mathbb{N}$. Thus, in the context of Theorem 11.3, we have independent $X_k = \mathbf{1}_{A_k}$ and

$$\mathbb{E}(W_n) = \sum_{k=1}^{n} \mathbb{E}(X_k) = \sum_{k=1}^{n} (1/k) \sim \log n\,.$$

We also know from Exercise 10.7 that we have the Strong Law of Large Numbers

$$\lim_{n\to\infty} \frac{W_n}{\mathbb{E}(W_n)} \;=\; \lim_{n\to\infty} \frac{W_n}{\log n} \;=\; 1\,, \qquad \mathbb{P}-\text{a.e.}$$

Now it is clear that $\text{Var}(W_n) = \sum_{k=1}^{n} \text{Var}(X_k) = \sum_{k=1}^{n} (k-1)/k^2 \;\sim\; \log n$. But we also have

$$\mathbb{E}\,|X_k - m_k|^3 \;=\; \big(1-(1/k)\big)^3 \cdot (1/k) + (1/k)^3 \cdot \big(1-(1/k)\big) \;\le\; (1/k) + (1/k)^3$$

thus $\sum_{j=1}^{n} \mathbb{E}\,\big|\, X_j - m_j \,\big|^3 \sim \log n$ and

$$s_n^{-3} \sum_{j=1}^{n} \mathbb{E}\,\big|\, X_j - m_j \,\big|^3 \;\le\; \log n \cdot (\log n)^{-3/2} \;\longrightarrow\; 0$$

as $n \to \infty$, so the condition (11.8) is satisfied and Theorem 11.3 thus vindicates our original guess, namely, that

$$\frac{W_n - \log n}{\sqrt{\log n}} \quad \text{converges in distribution to the standard Gaussian}\,.$$

It is often useful to state the LINDEBERG-FELLER Theorem directly in terms of (normalized) double arrays.

Theorem 11.4. LINDEBERG-FELLER CLT for Double Arrays: *Let $\{Y_{nj}\}_{1\le j\le k_n,\, n\in\mathbb{N}}$ be a double array of square-integrable and row-wise independent random variables; we denote by μ_{nj} their distributions, and assume*

$$\mathbb{E}(Y_{nj}) = \int x\,\mathrm{d}\mu_{nj}(x) = 0\,, \qquad \mathbb{E}(Y_{nj}^2) = \int x^2\,\mathrm{d}\mu_{nj}(x) =: \sigma_{nj}^2 > 0\,,$$

as well as $\sum_{j=1}^{k_n} \sigma_{nj}^2 = 1$ for every $n \in \mathbb{N}$. Then the following are equivalent:

(i) The Central Limit Theorem (CLT) *holds, namely,*

$$\mathfrak{Z}_n \;:=\; \sum_{j=1}^{k_n} Y_{nj} \;\; \textit{converges in distribution to the standard Gaussian} \tag{11.10}$$

as $n \to \infty$*; and so does the* FELLER condition

$$\lim_{n\to\infty} \left(\max_{1\le j\le k_n} \sigma_{nj}^2 \right) \,=\, 0\,. \tag{11.11}$$

(ii) The Central Limit Theorem (CLT) *holds; and the array is UAN.*
(iii) The LINDEBERG condition *holds:*

$$\lim_{n\to\infty} \sum_{j=1}^{k_n} \mathbb{E}\big(Y_{nj}^2\, \mathbf{1}_{\{|Y_{nj}|>\varepsilon\}}\big) \,=\, 0\,, \qquad \forall\ \varepsilon>0\,. \tag{11.12}$$

Furthermore, this latter condition is implied by the LYAPUNOV condition*: for some* $\delta > 0$*, we have*

$$\lim_{n\to\infty} \sum_{j=1}^{k_n} \mathbb{E}\big(\big|Y_{nj}\big|^{2+\delta}\big) \,=\, 0\,.$$

Example 11.2. Inversions in Random Permutations: Let us revisit the Example 7.7, when one deals with a *triangular* array $\{X_{nj}\}_{\substack{n\in\mathbb{N}\\ j=1,\cdots,n}}$ of row-wise independent random variables with

$$\mathbb{P}\big(X_{nj}=k\big) \,=\, \frac{1}{j}\,, \qquad k=0,1,\cdots,j-1\,, \quad j=1,\cdots,n\,,$$

thus

$$\mathbb{E}\big(X_{nj}\big) \,=\, \frac{j-1}{2}\,, \qquad \text{Var}\big(X_{nj}\big) \,=\, \frac{j^2-1}{12}\,, \qquad \mathbb{E}\big(X_{nj}^3\big) \,=\, \frac{j(j-1)^2}{4}\,.$$

The total number

$$S_n(\omega) \,=\, \sum_{j=1}^{n} X_{nj}(\omega)$$

of inversions in the permutation $\omega = (\omega_1,\cdots,\omega_n)$ of the integers $1,\cdots,n$, is thus a random variable with

$$\mathbb{E}\big(S_n\big) \,=\, \sum_{j=1}^{n} \frac{j-1}{2} \,\sim\, \frac{n^2}{4}\,, \qquad s_n^2 \,:=\, \text{Var}\big(S_n\big) \,=\, \sum_{j=1}^{n} \frac{j^2-1}{12} \,\sim\, \frac{n^3}{36}$$

so, for every $\varepsilon>0$, there exists an integer N_ε such that

$$\big|\, X_{nj}(\omega) - \mathbb{E}\big(X_{nj}\big)\,\big| \,\le\, j-1 \,\le\, n-1 \,\le\, \varepsilon s_n\,, \qquad \forall\ n\ge N_\varepsilon$$

holds for every permutation $\omega=(\omega_1,\cdots,\omega_n)$ of the integers $(1,\cdots,n)$.

Thus, we can apply Theorem 11.4 to the (normalized) triangular array $Y_{nj} = \big(X_{nj} - \mathbb{E}\big(X_{nj}\big)\big)/s_n$, $1\le j\le n$, $n\in\mathbb{N}$ to conclude that, as $n\to\infty$,

$$\lim_{n\to\infty}\ \mathbb{P}\left(\frac{S_n - (n^2/4)}{(n^{3/2})/6} \,\le\, x \right) \,=\, \Phi(x) := \int_{-\infty}^{x} \frac{e^{-\xi^2/2}}{\sqrt{2\pi}}\,\mathrm{d}\xi\,, \qquad \forall\, x\in\mathbb{R}\,.$$

To wit: among the $n!$ random permutations of the integers $(1,\cdots,n)$, there are roughly $\Phi(x)\,n!$ with no more than $(n^2/4) + x(n^{3/2})/6$ inversions, as $n\to\infty$.

This example underscores the necessity of considering double arrays in Central Limit Theory.

Exercise 11.1. Let $\xi_1, \xi_2, \xi_3, , \cdots$ be independent random variables with common BERNOULLI distribution $\mathbb{P}(\xi_n = 1) = \mathbb{P}(\xi_n = -1) = 1/2$. What can you say about the distribution of the random variable

$$Z_n = \frac{\xi_1 + 2\xi_2 + 3\xi_3 + \cdots + n\xi_n}{\sqrt{1 + 4 + 9 + \cdots + n^2}}, \qquad \text{as } n \to \infty\,?$$

Exercise 11.2. Let $X_1, X_2, X_3, , \cdots$ be independent random variables with common exponential distribution with parameter $\lambda > 0$, as in section 3.8. What can you say about the distribution of the random variable

$$Y_n = \lambda \big(\max_{1 \le j \le n} X_j \big) - \log n\,, \qquad \text{as } n \to \infty\,?$$

Exercise 11.3. Let $X_1, X_2, X_3, , \cdots$ be independent random variables with common "logistic" distribution $\mathbb{P}(X_j \le x) = F(x) = (1 + e^{-x})^{-1}$, $x \in \mathbb{R}$. What can you say about the distribution of the random variable

$$Y_n = \max_{1 \le j \le n} X_j - \log n\,, \qquad \text{as } n \to \infty\,?$$

Exercise 11.4. In the context of Example 7.7, show that the proportion of random permutations of the integers $(1, \cdots, n)$ that exhibit no more than $n^2/4$ inversions, approaches $1/2$ asymptotically as $n \to \infty$.

Exercise 11.5. Let $X_1, X_2, \cdots$ be independent random variables, with common distribution that has zero mean and variance equal to 1. With $S_n = \sum_{j=1}^n X_j$ show that, for $\mathbb{P}$-a.e. $\omega \in \Omega$, the set

$$\left\{ \frac{S_n(\omega)}{\sqrt{n}} \right\}_{n \in \mathbb{N}} \quad \text{is dense in } \mathbb{R}\,.$$

(*Hint:* Recall the HEWITT-SAVAGE 0-1 Law. We also have the following striking result

$$\lim_{n \to \infty} \frac{1}{\log n} \sum_{k=1}^{n} \frac{1}{k} \mathbf{1}_{\{a < S_k/\sqrt{k} \le b\}} = \int_a^b \frac{e^{-x^2/2}}{\sqrt{2\pi}}\, \mathrm{d}x\,, \qquad \mathbb{P} - \text{a.e.},$$

for any $-\infty \le a < b \le \infty$, an almost-sure version of the CLT; see, for instance, LACEY & PHILLIPP (1990) and the references there.)

11.2 Proof of the LINDEBERG-FELLER Central Limit Theorem

Let us adopt the setting and notation of Theorem 11.4, and suppose that the LINDEBERG condition (11.12) holds. We have already seen that this condition implies the requirement (11.11). We shall show here that it also implies the Central Limit Theorem assertion (11.10). In light of Lemma 11.1 and of Lemma 11.2, this will then prove the implications $(iii) \Rightarrow (i) \Rightarrow (ii)$; we shall not deal with the implication $(ii) \Rightarrow (iii)$.

We shall follow the treatment in BILLINGSLEY (1968), pp. 42-45, where we send the reader for the proof of the remaining implication in Theorem 11.4. This is essentially LINDEBERG's method, as presented by P. LÉVY (1925).

From Theorem 10.3 and Remark 10.1, it suffices to check

$$\lim_{n\to\infty} \mathbb{E}\big(f(\mathfrak{Z}_n)\big) \,=\, \mathbb{E}\big(f(Z)\big)\,, \quad \text{where } Z \sim \mathcal{N}(0,1) \text{ is standard Gaussian,} \tag{11.13}$$

for any given function $f : \mathbb{R} \to [0,1]$ of class $\mathcal{C}_b^\infty(\mathbb{R})$. The mean value theorem gives

$$g(z) \,:=\, \sup_{x\in\mathbb{R}} \Big(f(x+z) - f(x) - f'(x)\, z - \frac{1}{2}\, f''(x)\, z^2 \Big) \,\leq\, \kappa \left(z^2 \wedge |z|^3 \right) \tag{11.14}$$

for some real constant $\kappa > 0$; we also note

$$\Big| \big(f(x+y) - f(x+z)\big) - \Big(f'(x)\,(y-z) + \frac{1}{2}\, f''(x)\, \big(y^2 - z^2\big) \Big) \Big| \,\leq\, g(y) + g(z)\,. \tag{11.15}$$

If the variables in the double array $\mathcal{Y} = \big\{ Y_{nj} \big\}$ were all Gaussian, we would have equality in (11.13). Since they are not, we *introduce* – by enlarging the probability space, if necessary – an independent array $\mathcal{Z} = \big\{ Z_{nj} \big\}$ of row-wise independent, zero-mean Gaussian random variables with $\mathrm{Var}(Z_{nj}) = \sigma_{nj}^2$. In particular, each random variable Z_{nj} has the distribution of $\sigma_{nj}\, Z$.

We recall also from our discussion in section 7.4 that, for each $n \in \mathbb{N}$, the random variable $\sum_{j=1}^{k_n} Z_{nj}$ has Gaussian distribution with mean zero and variance $\sum_{j=1}^{k_n} \sigma_{nj}^2 = 1$; that is, exactly the distribution of the "target" random variable Z itself. [41]

The idea, then, is successively to replace the Y_{nj} by the Z_{nj}, namely, to consider

$$\begin{gathered}
\mathbb{E}\left(f(\,Y_{n1} + Y_{n2} + \cdots + Y_{n\,k_n-1} + Y_{n\,k_n})\right) \\
\mathbb{E}\left(f(\,Y_{n1} + Y_{n2} + \cdots + Y_{n\,k_n-1} + Z_{n\,k_n})\right) \\
\cdots\cdots \\
\mathbb{E}\left(f(\,Y_{n1} + Z_{n2} + \cdots + Z_{n\,k_n-1} + Z_{n\,k_n})\right) \\
\mathbb{E}\left(f(\,Z_{n1} + Z_{n2} + \cdots + Z_{n\,k_n-1} + Z_{n\,k_n})\right),
\end{gathered}$$

where the first term is $\mathbb{E}\big(f(\mathfrak{Z}_n)\big)$ and the last $\mathbb{E}\big(f(Z)\big)$. The plan now is to show that, for "n large, each successive term is so close to its predecessor, that the first and last terms are also very close".

[41] It is right at this point, that the properties of the Gaussian distribution, in particular its "infinite divisibility", are used in a crucial manner.

Thus, for each $j = 1, \cdots, k_n$ we consider the sum $W_{nj} := \sum_{i=1}^{j-1} Y_{ni} + \sum_{i=j+1}^{k_n} Z_{ni}$ (with an "empty" summation being interpreted as zero), and note

$$W_{n\,k_n} + Y_{n\,k_n} = \mathfrak{Z}_n\,, \qquad W_{n1} + Z_{n1} =: Z \sim \mathcal{N}(0,1)\,;$$

this gives the telescoping sum

$$f(\mathfrak{Z}_n) - f(Z) \;=\; \sum_{j=1}^{k_n} \big(f(W_{nj} + Y_{nj})\big) - f(W_{nj} + Z_{nj})\big)\,,$$

thus also the estimate

$$\big|\,\mathbb{E}\big(f(\mathfrak{Z}_n)\big) - \mathbb{E}\big(f(Z)\big)\,\big| \;\le\; \sum_{j=1}^{k_n} \big|\,\mathbb{E}\big(f(W_{nj} + Y_{nj})\big) - \mathbb{E}\big(f(W_{nj} + Z_{nj})\big)\,\big|\,. \tag{11.16}$$

The independence of the random variables W_{nj}, Y_{nj}, Z_{nj} leads to

$$\mathbb{E}\,\big(\,f'(W_{nj})\,(\,Y_{nj} - Z_{nj}\,)\big) \;=\; \mathbb{E}\,\big(\,f''(W_{nj})\,(\,Y_{nj}^2 - Z_{nj}^2\,)\big) \;=\; 0\,.$$

(This now explains why we took a second-order TAYLOR expansion in (11.14): the random variables Y_{nj}, Z_{nj} are known to have – the same – first and second moments; nothing is assumed regarding higher-order moments on the part of the Y_{nj}.) Therefore, in conjunction with (11.16), (11.15), we obtain

$$\big|\,\mathbb{E}\big(f(\mathfrak{Z}_n)\big) - \mathbb{E}\big(f(Z)\big)\,\big| \;\le\; \sum_{j=1}^{k_n} \mathbb{E}\big(g(Y_{nj})\big) \;+\; \sum_{j=1}^{k_n} \mathbb{E}\big(g(Z_{nj})\big)\,. \tag{11.17}$$

We need to show that each term on the right-hand side of (11.17) goes to zero as $n \to \infty$. For the first term, the identity $\sum_{j=1}^{k_n} \sigma_{nj}^2 = 1$ for every $n \in \mathbb{N}$ gives the estimate

$$\mathbb{E}\big(g(Y_{nj})\big) \;\le\; \kappa\,\Big(\int_{\{|Y_{nj}|<\varepsilon\}} |Y_{nj}|^3\,\mathrm{d}\mathbb{P} + \int_{\{|Y_{nj}|\ge\varepsilon\}} |Y_{nj}|^2\,\mathrm{d}\mathbb{P}\Big) \;\le\; \kappa\,\varepsilon\,\sigma_{nj}^2 \;+\; \kappa \int_{\{|Y_{nj}|\ge\varepsilon\}} |Y_{nj}|^2\,\mathrm{d}\mathbb{P}$$

which, in conjunction with (11.12), leads in the limit $n \to \infty$, then $\varepsilon \downarrow 0$, to

$$\sum_{j=1}^{k_n} \mathbb{E}\big(g(Y_{nj})\big) \;\le\; \kappa\,\varepsilon \;+\; \kappa \sum_{j=1}^{k_n} \int_{\{|Y_{nj}|\ge\varepsilon\}} |Y_{nj}|^2\,\mathrm{d}\mathbb{P} \;\longrightarrow\; 0\,.$$

For the second term in (11.17) we repeat the same steps, and note that we shall obtain

$$\lim_{n\to\infty} \sum_{j=1}^{k_n} \mathbb{E}\big(g(Z_{nj})\big) \;=\; 0$$

as soon as we have shown

$$\lim_{n\to\infty} \sum_{j=1}^{k_n} \int_{\{|Z_{nj}|\ge\varepsilon\}} |Z_{nj}|^2\,\mathrm{d}\mathbb{P} \;=\; 0$$

for all $\varepsilon \in (0,1)$. But thanks to (11.11) and $\sum_{j=1}^{k_n} \sigma_{nj}^2 = 1$, as well as of the fact that each random variable Z_{nj} has the distribution of $\sigma_{nj} Z$, we obtain

$$\sum_{j=1}^{k_n} \int_{\{|Z_{nj}|\geq\varepsilon\}} |Z_{nj}|^2 \, \mathrm{d}\mathbb{P} \;\leq\; \frac{1}{\varepsilon} \sum_{j=1}^{k_n} \mathbb{E}\big(|Z_{nj}|^3\big) \;\leq\; \frac{1}{\varepsilon}\, \mathbb{E}\big(|Z|^3\big) \sum_{j=1}^{k_n} \sigma_{nj}^3$$

$$\leq\; \frac{1}{\varepsilon}\, \mathbb{E}\big(|Z|^3\big) \Big(\max_{1\leq j\leq k_n} \sigma_{nj} \Big) \sum_{j=1}^{k_n} \sigma_{nj}^2 \;\longrightarrow\; 0$$

as $n \to \infty$, which concludes the proof of (11.10). □

PROOF OF THEOREM 11.1: With the setting of Theorem 11.1 we can cast

$$\frac{S_n - nm}{\sigma\sqrt{n}} = \sum_{j=1}^{k_n} Y_{nj} = \mathfrak{Z}_n\,, \qquad \text{where} \quad Y_{nj} := \frac{X_j - m}{\sigma\sqrt{n}}\,,\;\; j = 1, \cdots, k_n$$

and $k_n = n$, $\sigma_{nj}^2 = 1/n$. We are now in the setting of Theorem 11.4, and for every $\varepsilon \in (0,1)$ we have

$$\sum_{j=1}^{k_n} \int_{\{|Y_{nj}|\geq\varepsilon\}} |Y_{nj}|^2 \, \mathrm{d}\mathbb{P} \;=\; \frac{1}{\sigma^2} \int_{\{|X_1 - m|\geq\varepsilon\sigma\sqrt{n}\,\}} \big(X_1 - m\big)^2 \, \mathrm{d}\mathbb{P} \;\longrightarrow\; 0 \quad \text{as } n \to \infty$$

by dominated convergence. In other words, the LINDEBERG condition (11.12) is satisfied and therefore, as we just proved, (11.10) holds as well. □

11.3 Probabilistic Ideas in Arithmetic

In an amazing development, ERDÖS and KAC discovered in 1940 that prime divisors obey a Central Limit Theorem all their own.

Example 11.3. Prime Divisors: For any integer m, let us denote by $\xi(m)$ the number of prime divisors of the integer m (without multiplicities); for example, $\xi(5^6 \times 7^3) = 2$. Since there are infinitely-many primes, $\xi(m)$ is unbounded from above; but $\xi(m)$ drops back to 1 also for infinitely-many m, the primes and their powers. In other words, $\xi(m)$ varies in an irregular fashion, so it makes sense to ask what its average behavior is, and what the fluctuations around this average look like.

A century-old result of HARDY and RAMANUJAN (1920) states, roughly, that "*almost every* integer m has *approximately* $\log\log m$ prime divisors, and that the variance in the uncertainty of this statement is of the same order" $\log\log m$:

$$\lim_{n\to\infty} \frac{1}{n} \cdot \# \left\{ 1 \leq m \leq n \;:\; \frac{|\xi(m) - \log\log m|}{\sqrt{\log\log m}} \;>\; z_m \right\} \;=\; 0 \quad \text{and} \tag{11.18}$$

$$\lim_{n\to\infty} \frac{1}{n} \cdot \# \left\{ 1 \leq m \leq n \;:\; \frac{|\xi(m) - \log\log n|}{\sqrt{\log\log n}} \;>\; z_n \right\} \;=\; 0\,, \tag{11.19}$$

noindent for any sequence $\{z_n\}_{n\in\mathbb{N}}$ of positive numbers with $\lim_{n\to\infty} z_n = \infty$.

This statement was later refined in the classic paper by ERDÖS & KAC (1940), who showed, in the notation of (3.11), (3.12):

$$\lim_{n\to\infty} \frac{1}{n}\cdot\#\left\{ 1\le m\le n \ : \ x < \frac{\xi(m)-\log\log m}{\sqrt{\log\log m}} \le z \right\} = \Phi(z)-\Phi(x)\,, \quad \forall\, x<z\,, \tag{11.20}$$

$$\lim_{n\to\infty} \frac{1}{n}\cdot\#\left\{ 1\le m\le n \ : \ x < \frac{\xi(m)-\log\log n}{\sqrt{\log\log n}} \le z \right\} = \Phi(z)-\Phi(x)\,, \quad \forall\, x<z\,. \tag{11.21}$$

For example, for $-x = z = 0.9$ the above expression is about 0.6, which is thus the approximate proportion of m for which the ratio in (11.20) lies in the interval $[-.9,+.9]$. If m is near 10^{70}, this ratio is approximately $(\xi(m)-5)/\sqrt{5}$, which lies in $[-0.9,+0.9]$ if and only if $\xi(m)$ lies in $[5-0.9\sqrt{5}, 5+0.9\sqrt{5}] \approx [3,7]$. In other words, something close to 60 percent of the integers in the vicinity of 10^{70} have between 3 and 7 prime divisors. [42]

HARDY and RAMANUJAN came up with the statement (11.18)–(11.19), and gave a long and complicated proof for it. Then in (1934) P. TURÁN provided an elementary argument; this inspired his compatriot P. ERDÖS to interpret it probabilistically and, in collaboration with M. KAC, to formulate and prove the much sharper result (11.20) a few years later.

Intuitively, the ERDÖS-KAC result states that if n is a randomly chosen large integer, then the number of its distinct prime factors is approximately normally distributed, with mean and variance equal to $\log\log n$. KAC's idea here, is that "given a random natural number n, the events N_p expressing the fact that *the number n is divisible by some prime number p*, are independent for different primes"; and the sum $\sum_p \mathbf{1}_{N_p}$ counts how many distinct prime factors the integer n has. In this scheme of things, the idea is then to show that this sum satisfies the LINDEBERG condition, so the LINDEBERG form of the CLT, for independent but not identically distributed variables, guarantees that – appropriately centered and normalized – this sum will be standard Gaussian. The actual proof, due to ERDÖS, uses sieve theory to make the intuition rigorous.

The P. TURÁN argument for (11.19), expressed in current terminology, [43] runs as follows.

Proof of (11.19): One starts by introducing, for each $n\in\mathbb{N}$, a random variable M_n with *uniform distribution on* $\{1,\cdots,n\}$, namely

$$\mathbb{P}_n(M_n = m) = 1/n\,, \quad m = 1,\cdots,n\,.$$

We write the number of prime divisors of m as $\xi(m) = \sum_{p\le n}\eta_p(m)$, where the sum extends over primes and $\eta_p(m)=1$ if p/m, $\eta_p(m)=0$ otherwise, and note

$$\frac{1}{p}-\frac{1}{n} < \mathbb{E}_n\left[\eta_p(M_n)\right] = \frac{1}{n}\sum_{m=1}^{n}\eta_p(m) = \frac{1}{n}\lfloor n/p\rfloor \le \frac{1}{p}\,,$$

[42] A hollow sphere the size of the planet Earth filled with fine sand, would have around 10^{33} grains of sand. A volume the size of the observable universe would have around 10^{93} grains of sand.

[43] Back in 1934, P. TURÁN did not know that something called "the ČEBYŠEV inequality" existed, so he developed it himself.

as well as

$$\mathrm{Cov}_n\left(\eta_p(M_n),\eta_q(M_n)\right) = \mathbb{E}_n\left[\eta_p(M_n)\,\eta_q(M_n)\right] - \mathbb{E}_n\left[\eta_p(M_n)\right]\cdot\mathbb{E}\left[\eta_q(M_n)\right]$$

$$\leq \frac{1}{pq} - \left(\frac{1}{p}-\frac{1}{n}\right)\left(\frac{1}{q}-\frac{1}{n}\right) \leq \frac{1}{n}\left(\frac{1}{p}+\frac{1}{q}\right)$$

for any prime numbers $p \neq q$ (since then $\eta_p(m)\,\eta_q(m) = 1 \Leftrightarrow p|m$ and $q|m \Leftrightarrow \eta_{pq}(m) = 1$). Using the basic information

$$\pi(n) := \sum_{p\leq n} 1 \sim \frac{n}{\log n}\,, \qquad \sum_{p\leq n}\frac{1}{p} \sim \log\log n + A + O(1/\log n)$$

as $n \to \infty$ from the Prime-Number Theorem (e.g. APOSTOL (1976), Chapter 4), we obtain that the sum of covariances

$$\sum\sum\nolimits_{1\leq p\neq q\leq n}\mathrm{Cov}_n\big(\eta_p(M_n),\eta_q(M_n)\big) \leq \frac{1}{n}\cdot\sum\sum\nolimits_{p\neq q}\left(\frac{1}{p}+\frac{1}{q}\right) = \frac{\pi(n)-1}{n}\cdot\sum_{p\leq n}(2/p)$$

$$\sim 2\left(\frac{1}{\log n}-\frac{1}{n}\right)\left[\,\log\log n + A + O(1/\log n)\,\right] \longrightarrow 0$$

in the inequality of Exercise 4.9 is negligible as $n \to \infty$, yielding

$$\mathbb{E}_n(\xi(M_n)) = \sum_{p\leq n}\mathbb{E}_n\left[\eta_p(M_n)\right] = \log\log n + O(1) \qquad \text{and} \qquad \mathrm{Var}_n(\xi(M_n)) \leq \log\log n + O(1)\,.$$

It follows then from the ČEBYŠEV inequality that the left-hand side of (11.19) is asymptotically equivalent to

$$\mathbb{P}_n\left[\,|\xi(M_n) - \log\log n| > z_n\sqrt{\log\log n}\,\right] \leq \frac{\mathrm{Var}_n(\xi(M_n))}{z_n^2\,\log\log n} = \frac{1}{z_n^2} + O\big(\,1/\log\log n\,\big)$$

as $n \to \infty$, and the result (11.19) follows. □

Exercise 11.6. Deduce (11.18) from (11.19), using the very slow increase of $\log\log n$.
(*Hint:* Let $0 < \alpha < 1$, and consider only integers in the range $n^\alpha \leq m \leq n$; show that every integer m in this range that satisfies the condition of (5.19), also satisfies the condition of (11.18) for an appropriate increasing sequence $\{z_n\}_{n\in\mathbb{N}} \to \infty$.)

Exercise 11.7. Prove the Central-Limit-Theorem-type results (11.20)-(11.21) of ERDÖS and KAC. (*Hint:* Consult BILLINGSLEY (1986); DURRETT (2010), pages 133-137.)

Exercise 11.8. Argue that the ERDÖS-KAC result (11.20) implies the HARDY-RAMANUJAN result (11.18).

12 MARKOV Chains

Let $\mathfrak{X} = \{X_0, X_1, \cdots\}$ be a sequence of random variables, which are defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and take values in a countable set $\mathcal{S}$, the so-called *state space*.

We say that $\mathfrak{X}$ is a MARKOV *Chain* if, for all $n \in \mathbb{N}$ and all $(x_0, \cdots, x_{n+1}) \in \mathcal{S}^{n+2}$ for which $\mathbb{P}(X_0 = x_0, \cdots, X_n = x_n) > 0$, we have

$$\boxed{\mathbb{P}\big(X_{n+1} = x_{n+1} \,|\, X_0 = x_0, \cdots, X_{n-1} = x_{n-1}, X_n = x_n\big) \;=\; \mathbb{P}\big(X_{n+1} = x_{n+1} \,|\, X_n = x_n\big)\,.}$$

In somewhat loose but suggestive language, and recalling Exercise 7.1, this mandates that the "future" $A = \{X_{n+1} = x_{n+1}\}$ and the "past" $B = \{X_0 = x_0, \cdots, X_{n-1} = x_{n-1}\}$ be conditionally independent, given the "present" $C = \{X_n = x_n\}$.

If, in addition, we have

$$\mathbb{P}\big(X_{n+1} = y \,|\, X_n = x\big) \;=\; \mathbb{P}\big(X_1 = y \,|\, X_0 = x\big) \;=:\; p_{xy}\,, \quad \forall\ (x,y) \in \mathcal{S}^2\,,\ \ n \in \mathbb{N}\,,$$

then $\mathfrak{X}$ is called *time-homogeneous* MARKOV *Chain,* with transition probability matrix

$$\mathcal{P} \;:=\; \big(p_{xy}\big)_{(x,y)\in\mathcal{S}^2}\,.$$

This matrix contains nonnegative elements with the property

$$\sum_{y\in\mathcal{S}} p_{xy} \;=\; 1\,, \qquad \forall\, x \in \mathcal{S}\,;$$

such matrices are called *stochastic.*

We shall agree to write $\mathbb{P}_x(A)$ for $\mathbb{P}(A \,|\, X_0 = x)$, and set $\mathbb{P}_\pi(A) = \sum_{x\in\mathcal{S}} \pi(x)\, \mathbb{P}_x(A)$ for a probability measure π on the state-space. Integration with respect to the probability measure $\mathbb{P}_x$ ($\mathbb{P}_\pi$) will be denoted by $\mathbb{E}_x$ ($\mathbb{E}_\pi$).

We shall deal only with time homogeneous MARKOV Chains in this chapter. The simple random walk of Chapter 9 provides a salient example with $\mathcal{S} = \mathbb{Z}$ and $p_{x,x+1} = p$, $p_{x,x-1} = 1-p$, $\forall\, x \in \mathbb{Z}$ for some $p \in (0,1)$.

It may be useful to bear the following image in mind: think of X_n as the position on day $t = n$ of a particle, or of a "flea", that hops around the vertices of the grid (countable set) $\mathcal{S}$ visiting one site per day. The flea lives entirely in the present: it has neither memory (of what sites it has visited in the past), nor any sense of time (if it is today at site x, it will be tomorrow at site y with probability p_{xy}, irrespective of whether "today" is $n = 100$ or $n = 10^{35}$.

Important Remark: The so-called *initial distribution* μ of X_0, and the transition probability matrix $\mathcal{P}$, completely characterize the finite-dimensional distributions of $\mathfrak{X}$, since from the MARKOV property we have

$$\mathbb{P}\big(X_0 = x_0, X_1 = x_1, \cdots, X_n = x_n\big) \;=\; \mu(\{x_0\}) \cdot p_{x_0\,x_1}\, p_{x_1\,x_2}\, \cdots p_{x_{n-1}\,x_n} \tag{12.1}$$

for every $n \in \mathbb{N}$ and $(x_0, \cdots, x_n) \in \mathcal{S}^{n+1}$. These satisfy the consistency conditions of the DANIELL-KOLMOGOROV Theorem 6.2, so μ and $\mathcal{P}$ actually determine the distribution of $\mathfrak{X}$ viewed as a random element with values in $\mathcal{S}^{\mathbb{N}_0}$.

Theorem 12.1. MARKOV Property: *For any given* $N \in \mathbb{N}$ *and* $x \in \mathcal{S}$*, and conditioned on the event* $\{X_N = x\}$*, the sequence of random variables* $\{X_N,\, X_{N+1}, \cdots\}$ *is a* MARKOV *Chain and is independent of* $X_0, \cdots, X_{N-1}$*.*

This chain has transition probability matrix $\mathcal{P}$ *and initial distribution* δ_x*.*

In the terminology of ITÔ & MCKEAN (1974), this result expresses the property of "starting afresh" at each time $t = N$ for the MARKOV Chain. We shall present the proof of this result in Subsection 12.3 below; we shall also establish there a considerably stronger version of this result, the so-called Strong MARKOV Property, in Theorem 12.4.

Theorem 12.2. CHAPMAN-KOLMOGOROV Equations: *The* $n-$*step transition probabilities*

$$p_{xy}^{(n)} := \mathbb{P}\big(X_n = y \,|\, X_0 = x\big)\,, \qquad (x,y) \in \mathcal{S}^2\,,\ \ n \in \mathbb{N}$$

satisfy the CHAPMAN-KOLMOGOROV *equations*

$$\boxed{p_{xz}^{(m+n)} := \sum_{y \in \mathcal{S}} p_{xy}^{(m)}\, p_{yz}^{(n)}} \qquad \big((x,y) \in \mathcal{S}^2\,,\ \ (m,n) \in \mathbb{N}^2\big)\,.$$

In matrix notation, these equations amount to

$$\mathcal{P}^{(n)} := \big(p_{xy}^{(n)}\big)_{(x,y)\in\mathcal{S}^2} = \mathcal{P}^n \equiv \mathcal{P}\cdots\mathcal{P} \qquad (n - \textit{fold matrix product})\,,$$

and to

$$\mathcal{P}^{(m+n)} = \mathcal{P}^{(m)}\,\mathcal{P}^{(n)}\,.$$

Moreover, identifying the distribution of X_0 *as the row vector of weights* $\mu = \big(\mu(\{x\})\big)_{x\in\mathcal{S}}$*, we have*

$$\mathbb{P}(X_n = y) = \big(\mu\,\mathcal{P}^n\big)_y\,, \qquad \forall\ y \in \mathcal{S}\,.$$

The proof of this result is extremely simple: the particle just has to be somewhere on day $t = m$, so we list all possible alternatives, and use the law of total probability and the MARKOV property, to obtain

$$p_{xz}^{(m+n)} = \sum_{y\in\mathcal{S}} \mathbb{P}\big(X_{m+n} = z \,|\, X_0 = x,\, X_m = y\big)\cdot\mathbb{P}\big(X_m = y \,|\, X_0 = x\big) = \sum_{y\in\mathcal{S}} p_{xy}^{(m)}\, p_{yz}^{(n)}$$

and

$$\mathbb{P}(X_n = y) = \sum_{x\in\mathcal{S}} \mathbb{P}\big(X_n = y \,|\, X_0 = x\big)\cdot\mathbb{P}\big(X_0 = x\big) = \big(\mu\,\mathcal{P}^n\big)_y\,.$$

Illustration: With $\mathcal{S} = \{1,2\}$ and $0 < \alpha\,,\ \beta < 1$ consider

$$\mathcal{P} = \begin{pmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{pmatrix}.$$

We exploit the relation $\mathcal{P}^{n+1} = \mathcal{P}^n\,\mathcal{P}$, to write

$$p_{11}^{(n+1)} = p_{12}^{(n)}\,\beta + p_{11}^{(n)}\,(1-\alpha)\,, \qquad p_{11}^{(0)} = 1.$$

But $p_{11}^{(n)} + p_{12}^{(n)} = 1$, so we compute ($p_{11}^{(n)}$, and eventually the remaining entries) as

$$\mathcal{P}^n \,=\, \frac{1}{\alpha+\beta}\left(\begin{pmatrix}\beta & \alpha\\ \beta & \alpha\end{pmatrix} + \big(1-(\alpha+\beta)\big)^n \begin{pmatrix}\alpha & -\alpha\\ -\beta & \beta\end{pmatrix}\right) \longrightarrow \frac{1}{\alpha+\beta}\begin{pmatrix}\beta & \alpha\\ \beta & \alpha\end{pmatrix}$$

as $n \to \infty$. In particular, "in equilibrium" we have

$$\lim_{n\to\infty} \mathbb{P}_x\big(X_n = 1\big) \,=\, \frac{\beta}{\alpha+\beta}\,, \qquad \lim_{n\to\infty} \mathbb{P}_x\big(X_n = 2\big) \,=\, \frac{\alpha}{\alpha+\beta}\,.$$

Exercise 12.1. A Useful Characterization: A sequence $\mathfrak{X} = \{X_0,\, X_1, \cdots\}$ of random variables is a MARKOV chain with state space $\mathcal{S}$, transition probability matrix $\mathcal{P} = \big(p_{xy}\big)_{(x,y)\in\mathcal{S}^2}$ and initial distribution μ, if and only if (12.1) holds for every $n \in \mathbb{N}$ and $(x_0, \cdots, x_n) \in \mathcal{S}^{n+1}$.

Exercise 12.2. Consider a MARKOV Chain $\mathfrak{X} = \{X_0,\, X_1, \cdots\}$ with state space $\mathcal{S} = \{0, 1, 2\}$ and transition probability matrix

$$\mathcal{P} \,=\, \begin{pmatrix} 1 & 0 & 0\\ 1/4 & 1/2 & 1/4\\ 0 & 0 & 1\end{pmatrix}\,.$$

Detemine $\lim_{n\to\infty} \mathbb{P}_j(X_n = i)$ for $i, j = 0, 1, 2$.

Exercise 12.3. Suppose a virus can exist in N different strains, and that in each generation either it stays the same or – with probability $\alpha \in (0, 1)$ – mutates to another strain, chosen "at random" (that is, with equal probabilities among the remaining strains and independently of what has happened in previous generations).

What is the probability that, in the n^{th} generation, the strain is the same as it was at the beginning (that is, in the 0^{th} generation) ?

12.1 Hitting Times

Given a MARKOV Chain $\mathfrak{X} = \{X_0,\, X_1, \cdots\}$ with countable state space $\mathcal{S}$ and transition probability matrix $\mathcal{P}$, we fix a subset $A \subseteq \mathcal{S}$ of states and record the time

$$H^A \,:=\, \min\{\, n \in \mathbb{N}_0 \,|\, X_n \in A\,\}$$

it takes the chain to visit the set A. We shall employ throughout the convention $\min\emptyset = \infty$, so H^A is a random variable that takes values in $\mathbb{N}_0 \cup \{\infty\}$. We call this the *hitting time* of the set A.

With this setup, let us define the *hitting probabilities* and the expected hitting times

$$h^A(x) \,:=\, \mathbb{P}_x(H^A < \infty) \,=\, \mathbb{P}_x\big(X_n \in A\,,\ \text{for some } n \in \mathbb{N}_0\big)\,, \qquad k^A(x) \,:=\, \mathbb{E}_x(H^A)\,,$$

respectively, for $x \in \mathcal{S}$. These quantities can be characterized as follows.

Theorem 12.3. Smallest Nonnegative Solutions: *(i) The vector of hitting probabilities* $h^A = \{h^A(x)\}_{x\in\mathcal{S}}$ *is the minimal nonnegative solution to the system of linear equations*

$$h^A(x) = 1\,,\ \ x \in A \qquad \textit{and} \qquad h^A(x) = \sum_{y\in\mathcal{S}} p_{xy}\, h^A(y)\,,\ \ x \notin A\,.$$

Here the second equation expresses the "mean value", or "harmonicity", property.
(ii) The vector of expected hitting times $k^A = \{k^A(x)\}_{x\in\mathcal{S}}$ *is the minimal nonnegative solution to the system of linear equations*

$$k^A(x) = 0\,,\ \ x \in A \qquad \textit{and} \qquad k^A(x) = 1 + \sum_{y\notin A} p_{xy}\, k^A(y)\,,\ \ x \notin A\,.$$

Proof: (NORRIS (1997)) We shall use again and again the following method: "condition on the position of the particle at time $n = 1$, and use (7.3) or (7.6) along with the MARKOV property". (i) Indeed, with $x \notin A$, which guarantees $H^A \geq 1$, this gives the second equation for h^A straightaway:

$$h^A(x) := \mathbb{P}_x(H^A < \infty) = \sum_{y\in\mathcal{S}} \mathbb{P}_x(H^A < \infty \,|\, X_1 = y)\cdot \mathbb{P}_x(X_1 = y) = \sum_{y\in\mathcal{S}} p_{xy}\, h^A(y)\,,$$

since we have $\mathbb{P}_x(H^A < \infty \,|\, X_1 = y) = \mathbb{P}_y(H^A < \infty) = h^A(y)$ by the MARKOV property of Theorem 12.1.

Consider now an arbitrary vector $g = \{g(x)\}_{x\in\mathcal{S}}$ of nonnegative numbers that satisfy both $g(x) = 1\,,\ x \in A$ and $g(x) = \sum_{y\in\mathcal{S}} p_{xy}\, g(y)\,,\ x \notin A$; we need to show that $g(x) \geq h^A(x)$ holds for all $x \in \mathcal{S}$. Now, for $x \in A$ there is nothing to prove, as $g(x) = 1 \geq h^A(x)$. For $x \notin A$ we have clearly

$$g(x) = \sum_{y\in\mathcal{S}} p_{xy}\, g(y) = \sum_{y\in A} p_{xy} + \sum_{y\notin A} p_{xy}\, g(y) = \sum_{y\in A} p_{xy} + \sum_{y\notin A} p_{xy}\,\Big(\sum_{z\in A} p_{yz} + \sum_{z\notin A} p_{yz}\, g(z)\Big)$$

$$= \mathbb{P}_x(X_1 \in A) + \mathbb{P}_x(X_1 \notin A,\, X_2 \in A) + \sum_{y\notin A}\sum_{z\notin A} p_{xy}\, p_{yz}\, g(z)\,.$$

We keep going, to get

$$g(x) = \mathbb{P}_x(X_1 \in A) + \mathbb{P}_x(X_1 \notin A,\, X_2 \in A) + \cdots + \mathbb{P}_x(X_1 \notin A, \cdots, X_{n-1} \notin A,\, X_n \in A)$$

$$+ \sum_{y_1\notin A} \cdots \sum_{y_n\notin A} p_{xy_1}\, p_{y_1y_2} \cdots p_{y_{n-1}y_n}\, g(y_n)\,.$$

Now this last multiple sum is nonnegative; whereas the summation of all the other terms on the right-hand side is just $\mathbb{P}_x(H^A \leq n)$, the probability that the set A gets visited by day $t = n$. This means $g(x) \geq \mathbb{P}_x(H^A \leq n)$ for all $n \in \mathbb{N}$; thus, for every $x \in \mathcal{S}$ we obtain

$$g(x) \geq \lim_{n\to\infty} \mathbb{P}_x(H^A \leq n) = \mathbb{P}_x(H^A < \infty) = h^A(x)\,,$$

the minimality property claimed in part (i) the theorem. □

Exercise 12.4. Prove the second claim of Theorem 12.3.

Example 12.1. Gambler's Ruin Paradox: Consider the MARKOV Chain with $\mathcal{S} = \mathbb{N}_0$ and $p_{00} = 1$, $p_{x,x+1} = p$ and $p_{x,x-1} = q$ for $x \in \mathbb{N}$, where $0 < p = 1 - q < 1$. This is a simple random walk on the nonnegative integers with an absorbing barrier, at the origin.

You enter, in other words, a casino with $\$\,x$ in your pocket and gamble, \$1 at a time; with probability p you double your stake, with probability q you lose it. Let us assume the resources of the casino are limitless, so there is no upper bound to the fortune you can make. But what is the probability that you leave penniless? A version of this problem was first proposed by the Dutch mathematician Christiaan HUYGENS, who was visiting in Paris at the time, to Pierre DE FERMAT in 1657 – who provided a solution.

In our notation, let us compute $h(x) \equiv h^A(x)$ with $A = \{0\}$. We have $h(0) = 1$, as well as the linear, second-order difference equation

$$h(x) \;=\; p\,h(x+1) + q\,h(x-1)\,, \qquad x \in \mathbb{N}\,.$$

- For $p \neq 1/2$, the general solution of this difference equation is

$$h(x) \;=\; A + B\,\big(q/p\big)^x\,.$$

(i) If $p < 1/2$ the odds are against you; then the extra requirement $0 \le h(x) \le 1$ forces $B = 0$, so $h(x) = 1$ holds for all $x \in \mathbb{N}_0$ (because it does for $x = 0$, and $h(\cdot)$ is constant).
(ii) If $p > 1/2$ the odds are in your favor; since $h(0) = 1$, we get a family of solutions

$$h(x) \;=\; \big(q/p\big)^x \;+\; A\Big(1 - \big(q/p\big)^x\Big)\,;$$

for a nonnegative solution we need $A \ge 0$, so the minimal nonnegative solution is $h(x) = \big(q/p\big)^x$.

- For $p = 1/2$ the general solution of the equation is $h(x) = A + B\,x$ and again the requirement $0 \le h(x) \le 1$ forces $B = 0$, so $h(x) = 1$ for all $x \in \mathbb{N}_0$. Even if you play a fair game against an opponent with huge resources, you are certain eventually to end up broke. [44] □

12.2 Class Structures

Definition 12.1. Communication: *We say that a state $i \in \mathcal{S}$* leads to *state $j \in \mathcal{S}$ (or equivalently that state j is* accessible from *state i), and write $i \to j$, if*

$$\mathbb{P}_i\big(H^{\{j\}} < \infty\big) \;=\; \mathbb{P}_i\big(X_n = j\,,\ \textit{for some } n \in \mathbb{N}_0\big) \;>\; 0\,.$$

We say that state $i \in \mathcal{S}$ communicates with *state $j \in \mathcal{S}$, and write $i \leftrightarrow j$, if both $i \to j$ and $j \to i$ hold.*

[44] A lot of people find this disconcerting, whence the appellation of "paradox". Others, of the persuasion that "in life you have to keep working your way up, and that just trying to maintain the status quo leads you to decline", find it somewhat less disconcerting.

The relation $i \leftrightarrow j$ is reflexive ($i \leftrightarrow i$), symmetric ($i \leftrightarrow j$ is equivalent to $j \leftrightarrow i$) and transitive ($i \leftrightarrow j$ and $j \leftrightarrow k$ imply $i \leftrightarrow k$), thus an *equivalence relation* on $\mathcal{S}$.

As such, it partitions the state space into equivalence classes called *communication classes*.

Definition 12.2. Irreducibility, Closure: *(i) A Markov chain is called* irreducible, *if the entire state space $\mathcal{S}$ is a single communicating class.*
(ii) An equivalence class C is called closed, *if*

$$i \in C\,, \quad i \to j \quad \textit{imply} \quad j \in C\,;$$

once the particle finds itself in such a class, it is trapped there.
(iii) We call a state $i \in S$ absorbing, *if $\{i\}$ is a closed class.*

Exercise 12.5. A desperate gambler has only \$1, but needs \$5. He plays a (fair) coin-tossing game, and adopts the following strategy: on any given toss he wagers all he has, or the amount necessary to reach his goal, whichever of the two is smaller. This way, his "transition probabilities" are $p_{00} = p_{55} = 1$ and

$$p_{12} \,=\, p_{10} \,=\, p_{24} \,=\, p_{20} \,=\, p_{35} \,=\, p_{31} \,=\, p_{45} \,=\, p_{43} \,=\, 1/2\,.$$

What is the expected value of the time it will take him, starting with \$1, either to reach his goal or to lose everything?

Exercise 12.6. Consider a simple random walk $S_n \,=\, S_0 + \sum_{j=1}^{n} X_j$, $n \,\in\, \mathbb{N}$ and its current maximum $M_n := \max\{S_0,\, S_1,\, \cdots,\, S_n\}$ for $n \in \mathbb{N}_0$. Prove or disprove:
(i) $\{M_n\}_{n\in\mathbb{N}_0}$ is a MARKOV chain;
(ii) $\{(S_n, M_n)\}_{n\in\mathbb{N}_0}$ is a MARKOV chain;
(iii) $\{|S_n|\}_{n\in\mathbb{N}_0}$ is a MARKOV chain.

Exercise 12.7. Consider a MARKOV chain $\mathfrak{X} = \{X_n\}_{n\in\mathbb{N}_0}$ with state space partitioned as $\mathcal{S} = \cup A_k$ into (at most) countably many disjoint sets. Let $\mathfrak{Y} = \{Y_n\}_{n\in\mathbb{N}_0}$ be a process that takes the value y_k whenever the chain $\mathfrak{X}$ lies in the set A_k. Show that $\mathfrak{Y}$ is also a MARKOV chain, provided $p_{i_1 j} = p_{i_2 j}$ whenever i_1 and i_2 are in the same set A_k.

Exercise 12.8. For states $i \neq j$ in $\mathcal{S}$, the following are equivalent:
(i) $i \to j$;
(ii) $p_{ij}^{(n)} > 0$, for some $n \in \mathbb{N}_0$;
(iii) there exist $n \in \mathbb{N}_0$ and states $i_1, \cdots, i_n$ such that $p_{i i_1} p_{i_1 i_2} p_{i_2 i_3} \cdots p_{i_n j} > 0$.

Exercise 12.9. Birth-and-Death Chain: Consider the random walk on the nonnegative integers $\mathcal{S} = \mathbb{N}_0$ with transition probabilities

$$p_{i,i+1} = p_i \in (0,1)\,, \qquad p_{i,i-1} = q_i = 1 - p_i \qquad \text{for} \quad i \in \mathbb{N}$$

and $p_{0,0} = 1$ (absorbing state).

We start the resulting MARKOV chain at $X_0 \,=\, x \,\in\, \mathbb{N}$ (a population with x individuals). What is the probability $h(x)$ that the chain gets eventually absorbed at the origin; that is, that the population becomes extinct?

12.3 The Strong MARKOV Property

Let us consider again a MARKOV Chain $\mathfrak{X} = \{X_0,\, X_1, \cdots\}$ with countable state space $\mathcal{S}$ and transition probability matrix $\mathcal{P}$. For any $n \in \mathbb{N}_0$, we shall denote by

$$\mathcal{F}_n \,:=\, \sigma(X_0, \cdots, X_n)$$

the σ-algebra of all events that are associated with this MARKOV chain and observable up to time $t = n$.

Let us observe that $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \cdots \subseteq \mathcal{F}_{n-1} \subseteq \mathcal{F}_n \cdots$. We express this by saying that the family $\mathbb{F} = \{\mathcal{F}_n\}_{n\in\mathbb{N}_0}$ is a *filtration,* that is, an increasing sequence of sub-σ-algebras of $\mathcal{F}$. Let us also note that

$$A \in \mathcal{F}_n \quad \Longleftrightarrow \quad A \,=\, \bigcup_{(x_0,\cdots,x_n)\in C} \big\{X_0 = x_0, \cdots, X_n = x_n\big\} \quad \text{for some set } \; C \subseteq \mathcal{S}^{n+1}\,;$$

equivalently,

$$A \in \mathcal{F}_n \quad \Longleftrightarrow \quad \mathbf{1}_A = f_n(X_0, \cdots, X_n) \quad \text{for some function } \; f_n : \mathcal{S}^{n+1} \to \{0,1\}\,.$$

We express this by saying that the event $A \in \mathcal{F}_n$ "is determined by the random variables $X_0, \cdots, X_n$".

Proof of Theorem 12.1: For every event $A \in \mathcal{F} = \sigma(X_0, \cdots, X_N)$ (that is, determined by the random variables $X_0, \cdots, X_N$), we have to show for arbitrary $n \in \mathbb{N}$, $x \in \mathcal{S}$ and $(x_N, \cdots, x_{N+n}) \in \mathcal{S}^n$ the equality

$$\begin{aligned}\mathbb{P}\big(\{X_N = x_N, \cdots, X_{n+N} = x_{n+N}\} \cap A \,\big|\, X_N = x\big) \\ = \,\mathbf{1}_{x=x_N} \cdot p_{x_N x_{N+1}} \cdots p_{x_{N+n-1}x_{N+n}} \cdot \mathbb{P}(A \,|\, X_N = x)\end{aligned}$$

or equivalently

$$\begin{aligned}\mathbb{P}\big(\{X_N = x_N, \cdots, X_{n+N} = x_{n+N}\} \cap A \cap \{X_N = x\}\big) \\ = \,\mathbf{1}_{x=x_N} \cdot p_{x_N x_{N+1}} \cdots p_{x_{N+n-1}x_{N+n}} \cdot \mathbb{P}(A \cap \{X_N = x\})\,;\end{aligned} \tag{12.2}$$

then the result follows from Exercise 12.1. [45]

Now with $A = \{X_0 = x_0, \cdots, X_N = x_N\}$ for some $(x_0, \cdots, x_N) \in \mathcal{S}^{N+1}$, this equation holds trivially when $x_N \neq x$, as both its sides are zero; whereas when $x_N = x$, it amounts to

$$\mathbb{P}(X_0 = x_0, \cdots, X_{N+n} = x_{N+n}) \,=\, \mathbb{P}(X_0 = x_0, \cdots, X_N = x_N) \cdot p_{x_N x_{N+1}} \cdots p_{x_{N+n-1}x_{N+n}}$$

[45] For then, with $B := \{X_N = x_N, \cdots, X_{n+N} = x_{n+N}\}$, we have

$$\mathbb{P}\big(B \cap A \,\big|\, X_N = x\big) = \mathbf{1}_{x=x_N} \cdot p_{x_N x_{N+1}} \cdots p_{x_{N+n-1}x_{N+n}} \cdot \mathbb{P}(A \,|\, X_N = x)$$

for every $A \in \mathcal{F}_N$; therefore, by taking $A = \Omega$, we deduce $\mathbb{P}\big(B \,\big|\, X_N = x\big) = \mathbf{1}_{x=x_N} \cdot p_{x_N x_{N+1}} \cdots p_{x_{N+n-1}x_{N+n}}$. Thus, we have also the conditional independence $\mathbb{P}\big(B \cap A \,\big|\, X_N = x\big) = \mathbb{P}\big(B \,\big|\, X_N = x\big) \cdot \mathbb{P}\big(A \,\big|\, X_N = x\big)$ for every $A \in \mathcal{F}_N$: i.e., given the "present" $\{X_N = x\}$, the "future" B and the "past" A are conditionally independent.

which holds once again on the strength of Exercise 12.1.

In general, every set $A \in \mathcal{F}_n$ can be expressed as a countable union $A = \bigcup_{k\in\mathbb{N}} A_k$ of disjoint sets of the above form; thus, the result follows by adding up the resulting equalities for the sets $A_k\,,\ k \in \mathbb{N}\,.$ □

Definition 12.3. Stopping Time: *A random variable* $T : \Omega \to \mathbb{N}_0 \cup \{+\infty\}$ *is called* Stopping Time [46] *for the* MARKOV *Chain, if for every* $n \in \mathbb{N}_0$ *we have*

$$\{T = n\} \in \mathcal{F}_n\,, \quad \textit{or equivalently} \quad \{T \le n\} \in \mathcal{F}_n\,.$$

In other words: when T occurs, you know it straightaway. The primary examples of stopping times are the first hitting times H^A of the previous section: when the particle enters a set for the first time, you know it by observing its trajectory up to that time, since

$$\{H^A = n\} \;=\; \big\{X_0 \notin A, \cdots, X_{n-1} \notin A,\, X_n \in A\big\} \in \mathcal{F}_n\,.$$

Contrast this to the *last visitation time*

$$D^A \;:=\; \max\{\, n \in \mathbb{N}_0 \,|\, X_n \in A\,\} \qquad \text{(with the convention } \max\emptyset = 0)$$

into a given set A. This is typically not a stopping time, since

$$\{D^A = n\} \;=\; \big\{X_n \in A,\, X_{n+1} \notin A,\, X_{n+2} \notin A,\, \cdots\big\} \notin \mathcal{F}_n\,;$$

i.e., you need to "be a prophet", that is, have access to the entire future of the chain, to determine whether the event $\{D^A = n\}$ has occurred, or not.

On the other hand, degenerate random times, of the form $T = m$ for some $m \in \mathbb{N}_0$ (i.e., fixed dates), are also stopping times.

It is fairly easy to verify that, if T, $\{T_n\}_{n\in\mathbb{N}}$ and S are stopping times, so are

$$T+S\,,\ T\wedge S := \min\big(T,S\big)\,,\ T\vee S := \max\big(T,S\big)\,,\ \sup_{n\in\mathbb{N}} T_n\,,\ \inf_{n\in\mathbb{N}} T_n\,,\ \overline{\lim}_{n\in\mathbb{N}} T_n\,,\ \underline{\lim}_{n\in\mathbb{N}} T_n\,.$$

However, $T-1$ is typically NOT a stopping time (argue this).

The following is a very important generalization of Theorem 12.1; it says, effectively, that the MARKOV chain "starts afresh" not only at fixed times $n \in \mathbb{N}$ but also at every stopping time T. For its statement, let us recall from Exercise 7.1 the notion of conditional independence.

Theorem 12.4. Strong MARKOV Property: *Let* T *be a stopping time. Then, for any state* $x \in \mathcal{S}$ *such that* $\mathbb{P}\big(\{T<\infty\}\cap\{X_T = x\}\big) > 0$*, and conditional on the event* $\{T<\infty\}\cap\{X_T = x\}\,,$ *the random sequence* $\widehat{\mathfrak{X}} = \{\widehat{X}_n\}_{n\in\mathbb{N}_0}$ *defined by*

$$\widehat{X}_n \;:=\; X_{T+n}\,, \qquad n \in \mathbb{N}_0$$

is a MARKOV *Chain with transition probability matrix* $\mathcal{P}$ *and initial distribution* δ_x*.*

Furthermore, the random sequences $\widehat{\mathfrak{X}}$ *and* $\big\{X_{T\wedge n}\big\}_{n\in\mathbb{N}_0}$*, are then conditionally independent, given the event* $\{T<\infty\}\cap\{X_T = x\}\,.$

Proof: Take an arbitrary $A \in \sigma\big(X_{T\wedge n}\,,\ n \in \mathbb{N}_0\big)$ and observe that for every $m \in \mathbb{N}_0$ we have

$$A \cap \{T = m\} \;\in\; \sigma\big(X_0,\, X_1, \cdots,\, X_m\big) \;\in\; \mathcal{F}_m\,.$$

[46] An earlier terminology for these random variables is "MARKOV Times", in honor of A.A. MARKOV who seems to have been the first to study them.

The MARKOV property (12.2) at time $t = m$ gives

$$\begin{aligned}
&\mathbb{P}\left[\,\left\{\widehat{X}_1 = x_1, \cdots, \widehat{X}_n = x_n\right\} \cap A \cap \{T = m\} \cap \left\{X_T = x\right\}\,\right] \\
&= \mathbb{P}\left[\,\left\{X_{m+1} = x_1, \cdots, X_{m+n} = x_n\right\} \cap A \cap \{T = m\} \cap \left\{X_m = x\right\}\,\right] \\
&= \mathbb{P}_x\big(X_1 = x_1, \cdots, X_n = x_n\big) \cdot \mathbb{P}\left[\,A \cap \{T = m\} \cap \left\{X_m = x\right\}\,\right] \\
&= \mathbb{P}_x\big(X_1 = x_1, \cdots, X_n = x_n\big) \cdot \mathbb{P}\left[\,A \cap \{T = m\} \cap \left\{X_T = x\right\}\,\right].
\end{aligned}$$

More precisely: the first equality follows by definition; the second equality by virtue of (12.2) with A replaced by $A \cap \{T = m\} \in \mathcal{F}_m$ and $N = m$; whereas the third is rather obvious.

Now we add up over $m \in \mathbb{N}_0$, to obtain

$$\begin{aligned}
&\mathbb{P}\left[\,\left\{\widehat{X}_1 = x_1, \cdots, \widehat{X}_n = x_n\right\} \cap A \cap \{T < \infty\} \cap \left\{X_T = x\right\}\,\right] \\
&\qquad = \mathbb{P}_x\big(X_1 = x_1, \cdots, X_n = x_n\big) \cdot \mathbb{P}\left[\,A \cap \{T < \infty\} \cap \left\{X_T = x\right\}\,\right],
\end{aligned}$$

then divide by $\mathbb{P}\big(T < \infty\} \cap \{X_T = x\}\big) > 0$. We get

$$\begin{aligned}
&\mathbb{P}\left[\,\left\{\widehat{X}_1 = x_1, \cdots, \widehat{X}_n = x_n\right\} \cap A \,\middle|\, \{T < \infty\} \cap \left\{X_T = x\right\}\,\right] \\
&\qquad = \mathbb{P}_x\big(X_1 = x_1, \cdots, X_n = x_n\big) \cdot \mathbb{P}\left[\,A \,\middle|\, \{T < \infty\} \cap \left\{X_T = x\right\}\,\right],
\end{aligned}$$

thus also

$$\begin{aligned}
\mathbb{P}\left[\,\left\{\widehat{X}_1 = x_1, \cdots, \widehat{X}_n = x_n\right\} \cap A \,|\, \{T < \infty\} \cap \left\{X_T = x\right\}\,\right] &= \mathbb{P}_x\big(X_1 = x_1, \cdots, X_n = x_n\big) \\
&= \mathbb{P}\left[\,\left\{\widehat{X}_1 = x_1, \cdots, \widehat{X}_n = x_n\right\} \,|\, \{T < \infty\} \cap \left\{X_T = x\right\}\,\right].
\end{aligned}$$

Here, the last equality follows by taking $A = \Omega$ in the next-to-last display.

The equality of the first and third terms in the last display, gives the claimed conditional independence. The equality of the second and third terms in the last display, gives the transition probabilities of the MARKOV chain $\widehat{\mathfrak{X}} = \{\widehat{X}_n\}_{n \in \mathbb{N}_0}$. □

Example 12.2. (NORRIS (1997), EXAMPLE 1.4.3.) Let us put this theorem to work, by computing with its help the distribution of the first hitting time H_0 to the origin. We set, for $x \in \mathbb{Z}$,

$$H_x := \min\{n \in \mathbb{N}_0 \,|\, S_n = x\}, \qquad T_x := \min\{n \in \mathbb{N} \,|\, S_n = x\} \tag{12.3}$$

for the simple random walk $S_n = S_0 + \sum_{j=1}^n X_j$, $n \in \mathbb{N}_0$. Here $X_1, X_2, \cdots$ are independent Bernoulli random variables with $\mathbb{P}(X_j = 1) = p$ and $\mathbb{P}(X_j = -1) = q$, $0 < p = 1 - q < 1$.

We would like to do this by computing the generating function

$$g(s) := \mathbb{E}_1\big(s^{H_0}\big) = \sum_{n \in \mathbb{N}_0} s^n \, \mathbb{P}_1\big(H_0 = n\big), \qquad 0 < s < 1.$$

Here, as always, we denote $\mathbb{P}_x(A) = \mathbb{P}(A|S_0 = x)$ and $\mathbb{E}_x(\xi) = \mathbb{E}(\xi|S_0 = x)$ for an event A and for a random variable ξ, respectively.

We follow the "tried and true" method: condition on what happens on day $t = 1$. This gives

$$g(s) \;=\; \mathbb{E}_1\big(s^{H_0}\big) \;=\; \mathbb{E}_1\big(s^{H_0} \,|\, S_1 = 2\big) \cdot p \;+\; \mathbb{E}_1\big(s^{H_0} \,|\, S_1 = 0\big) \cdot q\,.$$

On the event $\{S_1 = 0\}$ we have clearly $H_0 = 1$, $\mathbb{P}_1-$a.e. ("with probability one, under $\mathbb{P}_1$"): if you start at $x = 1$ and your first step is down, you reach the origin in exactly one day. Thus, the last term on the above display is $\mathbb{E}_1\big(s^{H_0} \,|\, S_1 = 0\big) \cdot q = qs$.

By the MARKOV property at time 1, conditional on the event $\{S_1 = 2\}$ we have $H_0 \;=\; 1+\widehat{H}_0$, where $\widehat{H}_0$ is the time taken after time 1 to get to the origin. This random time $\widehat{H}_0$ has, under $\mathbb{P}_1$ and conditional on $\{S_1 = 2\}$, the same distribution as H_0 has under $\mathbb{P}_2$ unconditionally, thus

$$g(s) \;=\; p\;\mathbb{E}_1\big(s^{1+\widehat{H}_0}\,|S_1 = 2\big) \;+\; qs \;=\; ps\;\mathbb{E}_2(s^{H_0}) \;+\; qs\,. \tag{12.4}$$

Now, by the "tried and true" method yet again:

$$\mathbb{E}_2(s^{H_0}) = \mathbb{E}_2(s^{H_1+H_0'} \,|\, H_1 < \infty)\,\mathbb{P}(H_1 < \infty) \;+\; \mathbb{E}_2(s^{H_1+H_0'} \,|\, H_1 = \infty)\,\mathbb{P}(H_1 = \infty)\,, \tag{12.5}$$

where

$$H_0' \;:=\; \min\{n \in \mathbb{N}_0 \,|\, S_{H_1+n} = 0\}$$

is the first time after H_1 that the origin is visited ("to reach the origin starting at 2, you first have to get to 1; if you succeed, then comes the second leg of the trip, getting from 1 to the origin; and by the strong MARKOV property, the paths of the walk during these two legs of the trip are independent").

We claim that the last term on the right-hand side of (12.5) vanishes: for either $\mathbb{P}(H_1 = \infty)$ is zero or, if it is positive, then $\mathbb{E}_2(s^{H_1+H_0'} \,|\, H_1 = \infty)$ is zero.

We also claim that, conditional on $\{H_1 < \infty\}$, the random variables H_1 and H_0' are independent under $\mathbb{P}_2$; again as we argued before, this is a consequence of the strong MARKOV property. Thus

$$\begin{aligned}\mathbb{E}_2(s^{H_0}) \;&=\; \mathbb{E}_2(s^{H_0'} \,|\, H_1 < \infty) \cdot \mathbb{E}_2(s^{H_1} \,|\, H_1 < \infty) \cdot \mathbb{P}(H_1 < \infty) \\ &=\; \mathbb{E}_2(s^{H_0'} \,|\, H_1 < \infty) \cdot \mathbb{E}_2\big(s^{H_1}\,\mathbf{1}_{\{H_1<\infty\}}\big) \;=\; \mathbb{E}_2(s^{H_0'} \,|\, H_1 < \infty) \cdot \mathbb{E}_2\big(s^{H_1}\big)\,.\end{aligned}$$

But from the strong MARKOV property once again, the distribution of H_0' under $\mathbb{P}_2(\,\cdot\,|H_1 < \infty)$ is the same as the distribution of H_0 under $\mathbb{P}_1$; which is the same as the distribution of H_1 under $\mathbb{P}_2$ ("reaching 0 starting from 1, is like reaching 1 starting from 2" by the spatial homogeneity of the walk). Therefore,

$$\mathbb{E}_2(s^{H_0}) \;=\; \mathbb{E}_1(s^{H_0}) \cdot \mathbb{E}_2(s^{H_1}) \;=\; \big(\mathbb{E}_1(s^{H_0})\big)^2 \;=\; \big(g(s)\big)^2\,.$$

Substituting back in (12.4), we conclude

$$g(s) = ps\,(g(s))^2 + qs\,.$$

This is a quadratic equation for $g(s)$; it is solved as $g(s) = \big(1 \pm \sqrt{1 - 4pqs^2}\,\big)\,/\,(2ps)$. But the function $g(\cdot)$ is continuous and takes values in [0,1], so we are forced to take the negative root at $s = 0$ and then stick with it for all $0 \le s < 1$, namely,

$$\mathbb{E}_1\big(s^{H_0}\big) \;=\; g(s) \;=\; \frac{1 - \sqrt{1 - 4pqs^2}}{2ps}\,.$$

And a bit more generally, these arguments can be re-deployed to show, by induction,

$$\mathbb{E}_k(s^{H_0}) \;=\; \big(g(s)\big)^k\,, \qquad k \in \mathbb{N}\,.$$

(One can also expand the square root $\sqrt{1+x}$ as a power series in x, to get the distribution of H_0 under $\mathbb{P}_1$, namely

$$\begin{aligned} g(s) \;&=\; \frac{1}{2ps}\left\{1 - \left(1 + \frac{1}{2}\big(-4pqs^2\big) + \frac{1}{2}\big(-\frac{1}{2}\big)(-4pqs^2)/2! + \cdots\right)\right\} \\ &=\; qs + pq^2s^3 + \cdots \\ &=\; s\,\mathbb{P}_1(H_0=1) + s^2\,\mathbb{P}_1(H_0=2) + s^3\,\mathbb{P}_1(H_0=3) + \cdots\,.) \end{aligned}$$

Letting $s \uparrow 1$ we obtain

$$\mathbb{P}_1(H_0 < \infty) \;=\; \sum_{k\in\mathbb{N}_0} \mathbb{P}_1(H_0 = k) \;=\; \lim_{s\uparrow 1} g(s) \;=\; \frac{1-\sqrt{1-4pq}}{2p}\,.$$

Thus $\mathbb{P}_1(H_0 < \infty) = 1$ for $p \le 1/2$ and $\mathbb{P}_1(H_0 < \infty) = q/p$ for $p > 1/2$, in accordance with Example 12.1.

It is also not hard to compute the expected time to hit the origin via $\mathbb{E}_1(H_0) = \lim_{s\uparrow 1} g'(s)$. Of course $\mathbb{E}_1(H_0) = \infty$ for $p > 1/2$; so we take $p \le 1/2$ and differentiate $g(s) = ps\,(g(s))^2 + qs$, then pass to the limit as $s \uparrow 1$ to obtain

$$g'(s) \;=\; \frac{p(g(s))^2 + q}{1 - 2psg(s)} \;\longrightarrow\; \frac{1}{1-2p}\,, \qquad \text{thus} \qquad \mathbb{E}_1(H_0) \;=\; \frac{1}{1-2p}\,, \quad k \in \mathbb{N}$$

and, a bit more generally,

$$\boxed{\mathbb{E}_k(H_0) \;=\; \frac{k}{1-2p}\,, \quad k \in \mathbb{N}\,.}$$

In particular, $\mathbb{E}_k(H_0) = \infty$ even for $p = 1/2$. Here is, then, yet another twist to the "Gambler's Ruin Paradox" of Example 12.1: if you are playing a fair game against an adversary with huge resources, you are certain to lead yourself to eventual ruin – but this may take an awfully long time to come about!

Exercise 12.10. In the notation of the preceding Example, show that the generating function of the first return to the origin is

$$\mathbb{E}_0\big(s^{T_0}\big) \;=\; 1 - \sqrt{1-4pqs^2}\,.$$

This is the generating function of the probability distribution we computed in (9.12). Argue that, in the case $p = 1/2$, this computation leads easily to $\mathbb{E}_0(T_0) = \infty$.

12.4 Recurrence and Transience

Let us consider once again a MARKOV Chain $\mathfrak{X} = \{X_0, X_1, \cdots\}$ with countable state space $\mathcal{S}$ and transition probability matrix $\mathcal{P}$. For a given state $i \in \mathcal{S}$ we consider the *first passage time*

$$T_i := \min\big\{n \geq 1 : X_n = i\big\} \tag{12.6}$$

and inductively, for $r = 0, 1, 2, \cdots$, an entire sequence of passage times

$$T_i^{(0)} := 0\,, \qquad T_i^{(1)} := T_i\,, \qquad T_i^{(r+1)} := \min\big\{n \geq T_i^{(r)} + 1 \,|\, X_n = i\big\}\,. \tag{12.7}$$

We shall call

$$S_i^{(r)} := \big(T_i^{(r)} - T_i^{(r-1)}\big) \cdot \mathbf{1}_{\{T_i^{(r-1)} < \infty\}} \tag{12.8}$$

the *length of the* r^{th} *excursion* between successive visits.

From the strong MARKOV property, we have then the following rather elementary observations.

Exercise 12.11. For $r = 2, 3, \cdots$ and conditional on the event $\{T_i^{(r-1)} < \infty\}$, the excursion length $S_i^{(r)}$ is independent of $\{X_k\,,\ k = 1, \cdots, T_i^{(r-1)}\}$ and we have

$$\mathbb{P}\big(S_i^{(r)} = n \,|\, T_i^{(r-1)} < \infty\big) \,=\, \mathbb{P}_i(T_i = n)\,, \qquad n \in \mathbb{N}_0\,,$$

thus also

$$\mathbb{P}\big(S_i^{(r)} < \infty \,|\, T_i^{(r-1)} < \infty\big) \,=\, \mathbb{P}_i(T_i < \infty)\,. \qquad \square$$

Let us consider also the number of visits to the state i by the chain $\mathfrak{X}$ during its entire lifetime,

$$\boxed{V_i := \sum_{n \in \mathbb{N}_0} \mathbf{1}_{\{X_n = i\}}\,,}$$

and note that its expectation is $\mathbb{E}_i(V_i) = \sum_{n \in \mathbb{N}_0} p_{ii}^{(n)}$.

Definition 12.4. Recurrence and Transience: We say that the state i is *recurrent,* if

$$\mathbb{P}_i(V_i = \infty) \,=\, \mathbb{P}_i(X_n = i\,,\ \text{i.o.}) \,=\, 1\,;$$

we say that it is *transient,* if $\mathbb{P}_i(X_n = i\,,\ \text{i.o.}) = 0$, or equivalntly

$$\mathbb{P}_i(V_i < \infty) \,=\, \mathbb{P}(X_n \neq i\,,\ \text{for all but finitely many } n \in \mathbb{N}\,) \,=\, 1\,.$$

Lemma 12.1. *For* $r \in \mathbb{N}$ *we have*

$$\mathbb{P}_i(V_i > r) \,=\, (f_i)^r\,, \qquad \textit{where} \quad f_i := \mathbb{P}_i(T_i < \infty)\,.$$

Proof: Clearly $\{V_i > r\} = \{T_i^{(r)} < \infty\}$ when $X_0 = i$. The result holds for $r = 1$, because $\mathbb{P}_i(V_i > 1) = \mathbb{P}_i(T_i < \infty) = f_i$; assuming it holds for $r \in \mathbb{N}$, the Strong MARKOV property of Exercise 12.11 shows that

$$\mathbb{P}_i(V_i > r+1) \,=\, \mathbb{P}_i(T_i^{(r+1)} < \infty) \,=\, \mathbb{P}_i(T_i^{(r)} < \infty \text{ and } S_i^{(r+1)} < \infty)$$

$$= \mathbb{P}_i(S_i^{(r+1)} < \infty \,|\, T_i^{(r)} < \infty) \cdot \mathbb{P}_i(T_i^{(r)} < \infty) \,=\, \mathbb{P}_i(S_i^{(r+1)} < \infty \,|\, T_i^{(r)} < \infty) \cdot \mathbb{P}_i(V_i > r)$$

is equal to $\mathbb{P}_i(T_i^{(r)} < \infty)\,(f_i)^r = (f_i)^{r+1}$; the result follows by induction. □

Theorem 12.5. Recurrence/Transience Dichotomy: *For any state $i \in \mathcal{S}$, either*
(i) $\mathbb{P}_i(T_i < \infty) = 1$, in which case i is recurrent and $\sum_{n\in\mathbb{N}_0} p_{ii}^{(n)} = \infty$; or
(ii) $\mathbb{P}_i(T_i < \infty) < 1$, in which case i is transient and $\sum_{n\in\mathbb{N}_0} p_{ii}^{(n)} < \infty$.
In particular, every state $i \in \mathcal{S}$ is either recurrent or transient.

Proof: If $f_i = \mathbb{P}_i(T_i < \infty) = 1$, then the Lemma gives $\mathbb{P}_i(V_i = \infty) = \lim_{r\to\infty} \mathbb{P}_i(V_i > r) = 1$; thus i is recurrent, and $\sum_{n\in\mathbb{N}_0} p_{ii}^{(n)} = \mathbb{E}_i(V_i) = \infty$.

If $f_i = \mathbb{P}_i(T_i < \infty) < 1$, then

$$\sum_{n\in\mathbb{N}_0} p_{ii}^{(n)} \,=\, \mathbb{E}_i(V_i) \,=\, \sum_{r\in\mathbb{N}} \mathbb{P}_i(V_i \geq r) \,=\, \sum_{r\in\mathbb{N}} (f_i)^r \,=\, \frac{f_i}{1-f_i} \,<\, \infty\,;$$

in particular $\mathbb{P}_i(V_i < \infty) = 1$ and i is transient. □

Exercise 12.12. G. PÓLYA: (i) Show that the simple random walk on the integer lattice $\mathbb{Z}$ (to wit, $p_{x,x+1} = p \in (0,1)$, $p_{x,x-1} = q = 1-p$) is recurrent, if and only if it is symmetric ($p = 1/2$).

(ii) Show that the simple, symmetric random walk on the two-dimensional integer lattice $\mathbb{Z}^2$ (to wit, $p_{x,y} = 1/4$ if $\|x-y\| = 1$, and $p_{x,y} = 0$ otherwise) is recurrent.

(ii) Show that the simple, symmetric random walk on the three-dimensional integer lattice $\mathbb{Z}^3$ (to wit, $p_{x,y} = 1/6$ if $\|x-y\| = 1$, and $p_{x,y} = 0$ otherwise) is transient.

Theorem 12.6. Class Distinctions: *Let C be a communicating class. Then we have the following:*
(i) All states in C are either recurrent or transient.
(ii) If C is recurrent (that is, all states in C are recurrent), then it is closed.
(iii) If C is closed and of finite cardinality, it is recurrent.

Note that a closed class of *infinite* cardinality can easily be transient (that is, all states in C can be transient); the prime example is the simple random walk on the integer lattice with $p \neq 1/2$.

It is also not very hard to spot closed classes just from the structure of the transition probability matrix $\mathcal{P}$, so the recurrence or transience of classes of finite cardinality is not hard to determine.

Proof: (i) If $i \in C$ and $j \in C$, we have $p_{ij}^{(n)} > 0$, $p_{ji}^{(m)} > 0$ for some integers $n \geq 0$, $m \geq 0$, therefore

$$p_{ii}^{(n+\ell+m)} \,\geq\, p_{ij}^{(n)}\, p_{jj}^{(\ell)}\, p_{jj}^{(m)}\,, \qquad \forall\ \ell \in \mathbb{N}_0\,,$$

Now if i is transient, Theorem 12.5 gives

$$\sum_{\ell\in\mathbb{N}_0} p_{jj}^{(\ell)} \;\le\; \big(\, p_{ij}^{(n)}\, p_{jj}^{(m)}\,\big)^{-1} \sum_{\ell\in\mathbb{N}_0} p_{ii}^{(n+\ell+m)} \;<\; \infty\,;$$

this same Theorem 12.5 gives now the transience of j. If, on the other hand, i is recurrent, then either all other states in $\mathcal{C}$ are also recurrent, in which there is nothing to prove, or there is at least one transient state (let's call it j); but then the same argument as above shows that i must also be transient, a contradiction.

(ii) Suppose C is recurrent, but not closed. Then for some $i \in C$, $j \notin C$ and $m \in \mathbb{N}$ we have $\mathbb{P}_i(X_m = j) > 0$, and we claim that this gives

$$\mathbb{P}_i\big(\{X_m = j\} \cap \{X_n = i,\ \text{i.o.}\}\big) \;=\; 0\,, \tag{12.9}$$

thus $\mathbb{P}_i(X_n = i,\ \text{i.o.}) < 1$. This means that i is not recurrent, a contradiction.

To justify the claim (12.9), suppose the probability in question is positive; then we have $\mathbb{P}_i(X_m = j,\, X_{m+k} = i) > 0$ for some $k \in \mathbb{N}_0$ (it is possible to go from i to j and then back to i), so $i \leftrightarrow j$. But this contradicts $i \in C$, $j \notin C$. [47]

(iii) For any given starting state $j \in C$, we claim there exists a state $i \in C$ such that

$$0 \;<\; \mathbb{P}_j(X_n = i,\ \text{i.o.}) \;=\; \mathbb{P}_j(X_n = i,\ \text{for some } n \in \mathbb{N}) \cdot \mathbb{P}_i(X_n = i,\ \text{i.o.}) \tag{12.10}$$

by the strong MARKOV property. This gives $\mathbb{P}_i(X_n = i,\ \text{i.o.}) > 0$, so that i is not transient; therefore, it is recurrent (the dichotomy of Theorem 12.5).

To justify the positivity claim in (12.10), suppose on the contrary that for every state $i \in C$ we have $\mathbb{P}_j(X_n = i,\ \text{i.o.}) = 0$, that is, $\mathbb{P}_j\big(\, X_n \neq i,\ \text{for all but finitely many } n \in \mathbb{N}_0\big) = 1$. Because C is assumed to have finitely many elements this means that, after a finite number of days, the chain has to leave C, that is,

$$\mathbb{P}_j\big(\, X_n \notin C\,,\ \text{for all but finitely many } n \in \mathbb{N}_0\big) \;=\; 1\,.$$

But C is assumed to be closed, so this is impossible. □

Theorem 12.7. (Y.S. CHOW) Recurrence is Contagious: [48] *Suppose that $i \in \mathcal{S}$ is recurrent, and that $i \to j$. Then we have*

(a) $\mathbb{P}_j(X_n = i\,,\ i.o.) \;=\; 1\,;$

(b) $\mathbb{P}_j(T_i < \infty) \;=\; 1\,;$

(c) $j \to i\,;$

(d) $j \in \mathcal{S}$ *is recurrent.*

In particular, every communicating class $\mathcal{C}$ which is recurrent, is also closed.

Finally, if the MARKOV *chain is irreducible and recurrent, then for every $i \in \mathcal{S}$ and an arbitrary initial distribution on X_0 we have $\mathbb{P}(T_i < \infty) = 1$.*

[47] A stronger result is proved in Theorem 12.7 below.

[48] I learned this result when I was a student here at Columbia, and was taking Stochastic Processes from our Professor Y.S. CHOW. He wrote it up, and proved it, on the blackboard during class. I have not been able to locate it anywhere in the literature, then or since, so I call it "CHOW's Theorem".

Proof: By assumption, $p_{ij}^{(m)} > 0$ holds for some $m \in \mathbb{N}_0$. Furthermore, $\mathbb{P}_i(X_n = i, \text{ i.o.}) = 1$, so the MARKOV property gives

$$\begin{aligned} p_{ij}^{(m)} \;&=\; \mathbb{P}_i(X_m = j) \;=\; \mathbb{P}_i\big(\{X_m = j\} \cap \{X_n = i, \text{ i.o.}\}\big) \\ &= \mathbb{P}_i\big(\{X_m = j\} \cap \{X_{m+k} = i, \text{ for infinitely many } k \in \mathbb{N}_0\}\big) \\ &= \mathbb{P}_i\big(X_m = j\big) \cdot \mathbb{P}_i\big(X_{m+k} = i, \text{ for infinitely many } k \in \mathbb{N}_0 \,|\, X_m = j\big) \\ &= p_{ij}^{(m)} \cdot \mathbb{P}_j\big(X_k = i, \text{ for infinitely many } k \in \mathbb{N}_0\big)\,. \end{aligned}$$

This implies $\mathbb{P}_j\big(X_k = i, \text{ for infinitely many } k \in \mathbb{N}_0\big) = 1$, as well as

$$\mathbb{P}_j(T_i < \infty) \;=\; \mathbb{P}_j\big(X_k = i, \text{ for some } k \in \mathbb{N}_0\big) \;\geq\; \mathbb{P}_j\big(X_k = i, \text{ for infinitely many } k \in \mathbb{N}_0\big) \;=\; 1\,,$$

thus $j \to i$. Since $i \to j$ by assumption, we get $i \leftrightarrow j$ (the two states belong to the same equivalence class); and since i is recurrent and communicates with j, then j is also recurrent.

Now if the entire state space $\mathcal{S}$ consists of a single recurrent class, and if μ is the distribution of X_0, we have by the MARKOV property

$$\begin{aligned} \mathbb{P}(T_i < \infty) \;&=\; \sum_{j \in \mathcal{S}} \mathbb{P}(T_i < \infty \,|\, X_1 = j) \cdot \mathbb{P}(X_0 = j) \\ &= \sum_{j \in \mathcal{S}} \mathbb{P}_j(T_i < \infty) \cdot \mu(\{j\}) \;=\; \sum_{j \in \mathcal{S}} 1 \cdot \mu(\{j\}) \;=\; 1\,. \qquad \square \end{aligned}$$

Definition 12.5. Positive and Null Recurrence: If i is a recurrent state, we say that it is *positive recurrent*, if $\mathbb{E}_i(T_i) < \infty$; otherwise, we say that it is *null recurrent*.

For simple, symmetric random walk, the origin (in fact, any fixed state) is recurrent; yet the time to return to the origin, though a.s. finite, has infinite expectation, so this recurrence is null.

12.5 Invariant Distributions

Let us call a row vector $\lambda = \{\lambda_j\,,\ j \in \mathcal{S}\}$ a *measure* on the countable state space $\mathcal{S}$, if all its components are nonnegative; and let us agree to call such a measure a *probability measure* (or "distribution"), if in addition we have $\sum_{j \in \mathcal{S}} \lambda_j = 1$.

Suppose we are given a stochastic matrix $\mathcal{P} = \{p_{ij}\}_{(i,j) \in \mathcal{S}^2}$. We say that a measure λ on $\mathcal{S}$ is *invariant* for $\mathcal{P}$, if it satisfies the "balance inflow" condition

$$\lambda\,\mathcal{P} \;=\; \lambda, \qquad \text{i.e.,} \qquad \sum_{i \in \mathcal{S}} p_{ij}\lambda_i = \lambda_j, \;\; \forall\; j \in \mathcal{S}\,;$$

in other words, if it is a row eigenvector of $\mathcal{P}$ with eigenvalue one. Quite clearly, an invariant measure λ also satisfies $\lambda\,\mathcal{P}^n \;=\; \lambda$ for every $n \in \mathbb{N}$.

Example 12.3. When the transition probability matrix $\mathcal{P}$ *doubly stochastic,* that is, we have both

$$\sum_{j\in\mathcal{S}} p_{ij} \,=\, 1\,, \quad \forall\ i\in\mathcal{S} \qquad \text{and} \qquad \sum_{i\in\mathcal{S}} p_{ij} \,=\, 1\,, \quad \forall\ j\in\mathcal{S}\,,$$

then the measure $\lambda_i = 1\,,\ \forall\, i\in\mathcal{S}$ is clearly invariant.

If in addition the state-space $\mathcal{S}$ is finite, then the uniform distribution (normalized counting probability measure) $\lambda_i = 1/|\mathcal{S}|\,,\ \forall\, i\in\mathcal{S}$ is invariant.

Example 12.4. We have already seen that $\lambda = \big(\beta/(\alpha+\beta),\, \alpha/(\alpha+\beta)\big)$ is an equilibrium distribution for the Markov chain with transition matrix

$$\mathcal{P} \,=\, \begin{pmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{pmatrix}\,, \qquad \text{for given constants} \qquad 0<\alpha<1\,, \quad 0<\beta<1\,.$$

It is easy to check that this is the only invariant probability measure for the chain.

An invariant measure is also called *stationary;* this is because if the sequence of random variables $\{X_0, X_1, \cdots\}$ is a MARKOV Chain with transition matrix $\mathcal{P}$ and initial distribution (probability measure) λ, then so is the sequence $\{X_m, X_{m+1}, \cdots\}$ for every $m\in\mathbb{N}_0$. Indeed, for each $j\in\mathcal{S}$ we have then

$$\mathbb{P}(X_1 = j) = \sum_{i\in\mathcal{S}} \mathbb{P}(X_1 = j \,|\, X_0 = i)\cdot\mathbb{P}(X_0 = i) = \sum_{i\in\mathcal{S}} p_{ij}\lambda_i = \lambda_j,$$

and thus by induction also $\mathbb{P}(X_m = j) = (\lambda\,\mathcal{P}^m)_j = \lambda_j$ for every $m\in\mathbb{N}_0$; the remaining claims follow from the MARKOV property.

An invariant measure is also called *equilibrium measure,* because of the following property.

Lemma 12.2. *Suppose that $\mathfrak{X} = \{X_0,\, X_1, \cdots\}$ is a* MARKOV *Chain with finite state space $\mathcal{S}$ and transition probability matrix $\mathcal{P}$, and that the limits*

$$\lim_{n\to\infty} p_{ij}^{(n)} \,=\, \pi_j\,, \quad \forall\ j\in\mathcal{S} \quad \textit{exist for some } \ i\in\mathcal{S}\,.$$

Then $\pi = \{\pi_j\,,\ j\in\mathcal{S}\}$ is an invariant probability measure for the chain.

Proof: Because $\mathcal{S}$ is finite we can interchange summations and limits, to obtain

$$\sum_{j\in\mathcal{S}} \pi_j = \sum_{j\in\mathcal{S}} \lim_{n\to\infty} p_{ij}^{(n)} = \lim_{n\to\infty} \sum_{j\in\mathcal{S}} p_{ij}^{(n)} = 1$$

as well as

$$\pi_j \,=\, \lim_{n\to\infty} p_{ij}^{(n+1)} \,=\, \lim_{n\to\infty} \sum_{k\in\mathcal{S}} p_{ik}^{(n)}\, p_{kj} \,=\, \sum_{k\in\mathcal{S}} \lim_{n\to\infty} p_{ik}^{(n)}\, p_{kj} \,=\, \sum_{k\in\mathcal{S}} \pi_k\, p_{kj}\,. \qquad \square$$

We observe now that the row sums of the stochastic matrix $\mathcal{P}$ are all equal to one, so the column vector $\eta = (1, \cdots, 1)'$ is a column eigenvector with eigenvalue equal to one: the "balance outflow" (or *harmonicity*) condition

$$\sum_{j \in \mathcal{S}} p_{ij} \eta_j = \eta_i \,, \qquad \forall\ i \in \mathcal{S}$$

holds for this η: i.e., $\mathcal{P}\eta = \eta$. Thus, for a *finite state space* $\mathcal{S}$, there *always* exists also a row eigenvector with eigenvalue one, that is, an invariant measure. We shall show below, using probabilistic reasoning, that this result holds in great generality, for an arbitrary countable state space:

"Every irreducible and recurrent stochastic matrix $\mathcal{P}$ has an essentially (that is, modulo scalar multipliers) *unique invariant measure."*

• Let us start our construction of the (essentially unique) invariant measure of an irreducible and recurrent stochastic matrix $\mathcal{P}$. This construction will be probabilistic; and as we mentioned already, it is of interest only when the state space $\mathcal{S}$ has infinite cardinality. It will be carried out in the two theorems that follow.

For a given, fixed state $k \in \mathcal{S}$ (an "anchor", or "base"), let us recall the notation of (12.6) for the first passage time, and consider the expected number of visits

$$\gamma_i^k \ := \ \mathbb{E}_k \sum_{n=0}^{T_k - 1} \mathbf{1}_{\{X_n = i\}} \ = \ \mathbb{E}_k \sum_{n=1}^{T_k} \mathbf{1}_{\{X_n = i\}} \,, \tag{12.11}$$

to state $i \in \mathcal{S}$ during an "excursion", or "loop", away from (that is, between successive visits to) the base $k \in \mathcal{S}$. Please note that $\gamma_k^k = 1$ and recall the notation $T_k = \min\{n \geq 1 \,:\, X_n = k\}$ from (12.6).

Exercise 12.13. Justify the last equality in (12.11).

Theorem 12.8. Existence of Invariant Measure: *Suppose the stochastic matrix $\mathcal{P}$ is irreducible. If it is also recurrent, then we have $\gamma_k^k = 1$ and $0 < \gamma_i^k < \infty$ for all $i \in \mathcal{S}$; and the row the vector $\gamma^k = (\gamma_i^k \,,\, i \in \mathcal{S})$ is an invariant measure, that is,*

$$\gamma^k \mathcal{P} \ = \ \gamma^k \,.$$

The converse to the last statement is false: an invariant measure may exist, while the chain is irreducible but transient: the simple symmetric random walk on the three-dimensional integer lattice $\mathbb{Z}^3$ is transient (Exercise 12.12), yet it admits the invariant measure $\pi_i = 1$ for all $i \in \mathbb{Z}^3$, because the transition probability matrix is then doubly stochastic.

Theorem 12.9. Minimality and Uniqueness: *Suppose the stochastic matrix $\mathcal{P}$ is irreducible, and that λ is an invariant measure with $\lambda_k = 1$; then $\lambda \geq \gamma^k$.*

If, in addition, $\mathcal{P}$ is recurrent, then $\lambda = \gamma^k$.

Thus, for an irreducible and *recurrent* chain, all invariant measures are scalar multiples of γ^k (and, of course, of each other).

Recurrence is essential in the second sentence of the above theorem: for instance, the asymmetric simple random walk on the integers with $p \neq 1/2$ admits, as we have seen in Example 12.1, a two-parameter family $\pi_i = A + B(p/q)^i$, $i \in \mathbb{Z}$ of invariant measures satisfying the balance inflow condition

$$\pi_i = p\,\pi_{i+1} + q\,\pi_{i-1}\,, \qquad i \in \mathbb{Z}\,,$$

so uniqueness up to scalar multiples cannot possibly hold.

The requirement $\lambda_k = 1$ is just a convenient normalization. When the state-space is infinite, it may just not be possible to impose the more "intuitive" normalization $\sum_{j\in\mathcal{S}} \lambda_j = 1$, as we may have $\sum_{j\in\mathcal{S}} \lambda_j = \infty$ for *every* invariant measure λ.

The prime example of such a situation is the simple, symmetric random walk where, by irreducibility and recurrence, all invariant measures are scalar multiples of π with $\pi_i = 1$, solution of the balance equation $\pi_1 = (1/2)\,\pi_{i+1} + (1/2)\,\pi_{i-1}$, $i \in \mathbb{Z}$.

Theorem 12.10. Positive Recurrence: *Suppose the stochastic matrix $\mathcal{P}$ is irreducible; then the following are equivalent:*

(1) every state is positive recurrent;

(2) some state is positive recurrent;

(3) an invariant probability measure π exists.

Under any of the above conditions the mean recurrence times are given as

$$m_k := \mathbb{E}_k(T_k) = \frac{1}{\pi_k}\,, \qquad k \in \mathcal{S}\,. \tag{12.12}$$

For instance, consider the transition probability matrix

$$\mathcal{P} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 1/2 & 1/2 \\ 1/2 & 0 & 1/2 \end{pmatrix}\,,$$

for which $\pi\,\mathcal{P} = \pi$ amounts to

$$\pi_1 = (1/2)\,\pi_3\,, \qquad \pi_2 = \pi_1 + (1/2)\,\pi_3\,, \qquad \pi_3 = (1/2)\,\pi_2 + (1/2)\,\pi_3\,.$$

This gives $\pi_1 = 1/5$, $\pi_2 = \pi_3 = 2/5$ and, from Theorem 12.10, the mean recurrence times $\mathbb{E}_1(T_1) = 5$, $\mathbb{E}_2(T_2) = \mathbb{E}_3(T_3) = 2.5$.

We learn from Theorem 12.10 that, *for an irreducible and recurrent chain, either all states are positive-recurrent, or all states are null-recurrent.*

The simple symmetric random walk on the integer lattice provides an example of a chain with a countable infinity of states, whose every state is null-recurrent; recall Exercise 12.10, and observe that every state communicates with every other state in this chain. For an example of a chain with a countable infinity of states, whose every state is positive recurrent, see Exercise 12.15.

Proof of Theorem 12.8: For every $n \in \mathbb{N}$, the event $\{T_k \geq n\}$ is determined entirely based on $X_0, X_1, \cdots, X_{n-1}$, that is $\{T_k \geq n\} = \{T_k > n-1\} = \{X_0 \neq k, X_1 \neq k, \cdots, X_{n-1} \neq k\}$. Thus, the MARKOV property gives

$$\begin{aligned}\mathbb{P}_k\big(X_{n-1} = i,\, X_n = j,\, T_k \geq n\big) &= \mathbb{P}_k\big(X_n = j \,\big|\, X_{n-1} = i\,,\, T_k \geq n\big) \cdot \mathbb{P}_k\big(X_{n-1} = i,\, T_k \geq n\big) \\ &= p_{ij} \cdot \mathbb{P}_k\big(X_{n-1} = i,\, T_k \geq n\big)\,;\end{aligned}$$

on the other hand, recurrence implies that we have both $T_k < \infty$ and $X_0 = X_{T_k} = k$ a.s. under $\mathbb{P}_k$, so the MARKOV property gives

$$\gamma_j^k = \mathbb{E}_k \sum_{n=1}^{T_k} \mathbf{1}_{\{X_n = j\}} = \mathbb{E}_k \sum_{n\in\mathbb{N}} \mathbf{1}_{\{X_n = j,\, T_k \geq n\}} = \sum_{n\in\mathbb{N}} \mathbb{P}_k\big(X_n = j,\, T_k \geq n\big) \tag{12.13}$$

$$= \sum_{i\in\mathcal{S}} \sum_{n\in\mathbb{N}} \mathbb{P}_k\big(X_{n-1} = i,\, X_n = j,\, T_k \geq n\big) = \sum_{i\in\mathcal{S}} p_{ij} \sum_{n\in\mathbb{N}} \mathbb{P}_k\big(X_{n-1} = i,\, T_k \geq n\big)$$

$$= \sum_{i\in\mathcal{S}} p_{ij}\; \mathbb{E}_k \sum_{m\in\mathbb{N}_0} \mathbf{1}_{\{X_m = i,\, T_k \geq m+1\}} = \sum_{i\in\mathcal{S}} p_{ij}\; \mathbb{E}_k \sum_{m=0}^{T_k - 1} \mathbf{1}_{\{X_m = i\}} = \sum_{i\in\mathcal{S}} p_{ij}\, \gamma_i^k\,.$$

This shows that $\gamma^k = (\gamma_i^k\,,\, i \in \mathcal{S})$ is an invariant measure.

Finally, since the chain is irreducible, there exist for each state $i \in \mathcal{S}$ integers m and n such that $p_{ik}^{(n)} > 0$ and $p_{ki}^{(m)} > 0$, so $\gamma_i^k \geq \gamma_k^k\, p_{ki}^{(m)} = p_{ki}^{(m)} > 0$ and $\gamma_i^k\, p_{ik}^{(n)} \leq \gamma_k^k = 1$. □

Proof of Theorem 12.9: For every state $j \in \mathcal{S}$ we have

$$\lambda_j = \sum_{i_0\in\mathcal{S}} \lambda_{i_0}\, p_{i_0 j} = \sum_{i_0\neq k} \lambda_{i_0}\, p_{i_0 j} + p_{kj} = \sum_{i_0\neq k} \sum_{i_1\neq k} \lambda_{i_1}\, p_{i_1 i_0}\, p_{i_0 j} + p_{kj} + \sum_{i_0\neq k} p_{k i_0}\, p_{i_0 j},$$

where the third equality is obtained from the first; and inductively,

$$\lambda_j = \sum_{i_0\neq k} \cdots \sum_{i_n\neq k} \lambda_{i_n}\, p_{i_n i_{n-1}} \cdots p_{i_0 j} + p_{kj} + \sum_{i_0\neq k} p_{k i_0}\, p_{i_0 j} + \cdots + \sum_{i_0\neq k} \cdots \sum_{i_n\neq k} p_{k i_{n-1}} \cdots p_{i_0 j}$$

$$\geq \mathbb{P}_k\big(X_1 = j,\, T_k \geq 1\big) + \mathbb{P}_k\big(X_2 = j,\, T_k \geq 2\big) + \cdots + \mathbb{P}_k\big(X_n = j,\, T_k \geq n\big) \longrightarrow \gamma_j^k$$

as $n \to \infty$, thanks to (12.13). This proves $\lambda \geq \gamma^k$.

If the chain is recurrent, then the measure γ^k is invariant (Theorem 12.8); thus $\mu := \lambda - \gamma^k$ is also invariant, as well as a measure (all its components are nonnegative). For every given $i \in \mathcal{S}$, irreducibility implies $p_{ij}^{(n)} > 0$ for some integer n, and this gives

$$0 = \lambda_k - \gamma_k^k = \mu_k = \sum_{j\in\mathcal{S}} \mu_j\, p_{jk}^{(n)} \geq \mu_i\, p_{ik}^{(n)} \geq 0\,, \qquad \text{forcing} \quad \mu_i = 0\,. \qquad \square$$

Proof of Theorem 12.10: The implication (1) → (2) is trivial.

To prove (2) → (3), suppose a given, fixed state $k \in \mathcal{S}$ is positive recurrent; then the entire chain consists of a single, recurrent equivalence class, and $\gamma_j^k = \mathbb{E}_k \sum_{n=1}^{T_k} \mathbf{1}_{\{X_n = j\}}\,,\ j \in \mathcal{S}$ is an invariant measure. But this gives

$$\sum_{j\in\mathcal{S}} \gamma_j^k = \mathbb{E}_k(T_k) = m_k < \infty\,, \qquad \text{therefore} \qquad \pi_j := \gamma_j^k / m_k\,, \quad j \in \mathcal{S}$$

defines then an invariant probability measure.

It remains to prove (3) $\to$ (1). Fix an arbitrary state $k \in \mathcal{S}$ and note that irreducibility and $\sum_{i\in\mathcal{S}} \pi_i = 1$

$$\pi_k \;=\; \sum_{i\in\mathcal{S}} \pi_i \, p_{ik}^{(n)} \,, \qquad \text{for } \; n \in \mathbb{N}_0$$

imply $\pi_k > 0$. (Argue this out in detail!) [49] Then $\lambda_j := \pi_j/\pi_k \,, \;\; j \in \mathcal{S}$ defines an invariant measure, and $\lambda_j \geq \gamma_j^k \,, \; \forall \, j \in \mathcal{S}$ follows from Theorem 12.9; thus

$$\mathbb{E}_k(T_k) \;=\; \sum_{j\in\mathcal{S}} \gamma_j^k \;\leq\; \sum_{j\in\mathcal{S}} \big(\pi_j/\pi_k\big) \;=\; 1/\pi_k \;<\; \infty \,, \tag{12.14}$$

so k is positive recurrent. Because the chain is irreducible, and the selected state k was arbitrary, every state is positive recurrent. But then we can deploy Theorem 12.9 once again, to obtain $\lambda_j = \gamma_j^k \,, \; \forall \, j \in \mathcal{S}$, and deduce that (12.14) holds as an equality. $\square$

Definition 12.6. Detailed Balance and Ergodicity: *A stochastic matrix $\mathcal{P}$ and a measure λ on the state space $\mathcal{S}$ are said to be in* Detailed Balance, *if*

$$\lambda_i \, p_{ij} \;=\; \lambda_j \, p_{ji} \,, \qquad \forall \;\; (i,j) \in \mathcal{S}^2 \,.$$

An invariant distribution π for the stochastic matrix $\mathcal{P}$ is called Ergodic, *if every bounded function $f : \mathcal{S} \to \mathbb{R}$ which satisfies the "harmonicity" condition $f = \mathcal{P} f$, is $\pi-$a.e. constant.*

MARKOV Chains whose invariant distributions satisfy the conditions of detailed balance and ergodicity, have some quite remarkable properties. We shall see some of them below.

Exercise 12.14. A professor owns $N \geq 1$ umbrellas. He walks to the office in the morning and walks home in the evening. If it is raining he likes to carry an umbrella, and if it is fine he does not. Suppose that it rains on each journey with probability $p \in (0,1)$, independently of past weather.

(i) Describe the evolution of the MARKOV chain

$$X_n \;=\; \text{number of umbrellas at home on morning of day } n \;\; (n = 0, 1, \cdots, N)$$

by a graph with labeled edges indicating the transition probabilities. (No explanatory arguments needed.)

(ii) Argue why this chain is irreducible, aperiodic, and positive recurrent.

(iii) Show that the unique invariant distribution satisfies $\pi_1 \;=\; \pi_2 \;=\; \cdots \;=\; \pi_N$, and determine π_0 and π_N.

(iv) What is the long-run proportion of journeys on which the professor gets wet? Give reasons for your answer.

[49] For suppose $\pi_k = 0$. Then the display gives $\pi_i \, p_{ik}^{(n)} = 0$ for all $n \in \mathbb{N}_0$, $i \in \mathcal{S}$. If $\pi_i = 0$ for all $i \in \mathcal{S}$, then we get the absurdity $\sum_{i\in\mathcal{S}} \pi_i = 0$. If on the other hand $\pi_i > 0$ holds for some $i \in \mathcal{S}$, the display gives $p_{ik}^{(n)} = 0$ for all $n \in \mathbb{N}_0$ which is impossible because then k is not accessible from i and irreducibility fails.

Exercise 12.15. Let $\mathfrak{X} = \{X_0,\, X_1, \cdots\}$ be a MARKOV Chain with state space $\mathcal{S} = \mathbb{N}_0$. For each of the following transition probability matrices $\mathcal{P}$, argue that the chain is irreducible and examine whether it is positive recurrent, null recurrent or transient.

(i) $p_{x,\,0} = 1/(x+2)$, $p_{x,\,x+1} = (x+1)/(x+2)$;

(ii) $p_{x,\,0} = (x+1)/(x+2)$, $p_{x,\,x+1} = 1/(x+2)$;

(iii) $p_{x,\,0} = 1/(x^2+2)$, $p_{x,\,x+1} = (x^2+1)/(x^2+2)$.

If the chain is positive recurrent, compute also the invariant probability measure.

12.6 Limit Theory for MARKOV Chains

The following two results, Theorems 12.11 and 12.12, are fundamental. The first deals with convergence of MARKOV Chains to equilibrium, whereas the second is an ergodic result, the analogue of the Strong Law of Large Numbers in the present context. The next result, the Central Limit Theorem 12.13 for MARKOV Chains, we just state; its proof is well beyond the scope of the machinery at our disposal.

In order to formulate the first two of these results, we shall need the following notion.

A state i will be called *aperiodic,* if $p_{ii}^{(n)} > 0$ holds for all sufficiently large $n \in \mathbb{N}$.

Exercise 12.16. A given state i is aperiodic, if and only if $\{n \in \mathbb{N} \,:\, p_{ii}^{(n)} > 0\}$ has no common divisor other than 1.

Exercise 12.17. Suppose a given MARKOV Chain is irreducible, and contains an aperiodic state i. Then all states are aperiodic.

Theorem 12.11. Convergence to Equilibrium: *Suppose that $\mathcal{P}$ is irreducible, aperiodic, and possesses an invariant distribution π. If $\mathfrak{X} = \{X_0,\, X_1, \cdots\}$ is a* MARKOV *Chain, with transition probability matrix $\mathcal{P}$ and with* arbitrary *initial distribution λ, we have*

$$\lim_{n\to\infty} \mathbb{P}\big(X_n = j\big) \,=\, \pi_j\,, \qquad \forall\;\; j \in \mathcal{S}\,.$$

In particular, $\lim_{n\to\infty} p_{ij}^{(n)} = \pi_j$ *for all* $(i,j) \in \mathcal{S}^2$.

A MARKOV Chain with state-space $\{1,2\}$ and transition probability matrix

$$\mathcal{P} \,=\, \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \tag{12.15}$$

satisfies $\mathcal{P}^{2n} = I$ and $\mathcal{P}^{2n+1} = \mathcal{P}$ for all $n \in \mathbb{N}$; this chain is thus periodic, and the $p_{ij}^{(n)}$ fail to converge as $n \to \infty$. On the other hand, the row vector $\pi = (1/2, 1/2)$ gives the unique invariant distribution for this chain.

Theorem 12.12. SLLN for MARKOV Chains: *Suppose* $\mathfrak{X} = \{X_0,\, X_1, \cdots\}$ *is a* MARKOV *Chain with irreducible transition probability matrix* $\mathcal{P}$ *and* arbitrary *initial distribution* λ. *Then in the notation of* (12.12) *we have, for any state* $k \in \mathcal{S}$, *the property*

$$\lim_{N\to\infty} \frac{1}{N} \sum_{t=0}^{N-1} \mathbf{1}_{\{X_t = k\}} \;=\; \frac{1}{m_k}\,, \qquad \textit{w.p. } 1. \tag{12.16}$$

If, furthermore, $\mathcal{P}$ *is irreducible and positive recurrent, then for any bounded* $f : \mathcal{S} \to \mathbb{R}$ *we have in the notation of* (12.12) *the Strong Law of Large Numbers*

$$\lim_{N\to\infty} \frac{1}{N} \sum_{t=0}^{N-1} f\big(X_t\big) \;=\; \sum_{k\in\mathcal{S}} \frac{f(k)}{m_k}\,, \qquad \textit{w.p. } 1. \tag{12.17}$$

Proof of Theorem 12.11: Consider an independent MARKOV Chain $\mathfrak{Y} = \{Y_0,\, Y_1, \cdots\}$ with the same transition probability matrix $\mathcal{P}$ but initial distribution π, the invariant probability distribution. Fix a reference state ("anchor", or "base") $\mathfrak{b} \in \mathcal{S}$ and consider the first time

$$T \;:=\; \min\big\{n \geq 1 \,:\, X_n = Y_n = \mathfrak{b}\,\big\}$$

that both chains find themselves there.

- We claim that $\mathbb{P}(T < \infty) = 1$.

To see this, observe that $\mathfrak{W} = \{W_0,\, W_1, \cdots\}$ with $W_n = (X_n, Y_n)$, is also a MARKOV Chain with state space $\mathcal{S} \times \mathcal{S}$, transition probabilities

$$\widetilde{p}_{(i,k)(j,\ell)} \;=\; p_{ij}\, p_{k\ell},$$

and initial and invariant probability measures given by

$$\mu_{(i,k)} = \lambda_i\, \pi_k \qquad \text{and} \qquad \widetilde{\pi}_{(i,k)} = \pi_i\, \pi_k\,,$$

respectively. Because the original chain is aperiodic, we have

$$\widetilde{p}^{\,(n)}_{(i,k),(j,\ell)} \;=\; p^{(n)}_{ij}\, p^{(n)}_{k\ell} > 0$$

for every two states (i,k) and (j,ℓ) in $\mathcal{S}\times\mathcal{S}$ and $n \in \mathbb{N}$ a sufficiently large integer, so the new, "product" chain $\mathfrak{W}$ is irreducible.

By Theorem 12.10, the chain $\mathfrak{W}$ is positive recurrent; and since T is the first passage time of $\mathfrak{W}$ to the state $(\mathfrak{b},\mathfrak{b})$, we conclude from Theorem 12.7 that $\mathbb{P}(T < \infty) = 1$.

- The idea now, is to *switch paths at time* T; that is, to consider the random sequence $\mathfrak{Z} = \{Z_0,\, Z_1, \cdots\}$ with

$$Z_n \;:=\; X_n\,, \;\text{ for } n < T \qquad \text{and} \qquad Z_n \;:=\; Y_n\,, \;\text{ for } n \geq T\,.$$

Using the strong MARKOV property and the independence of $\mathfrak{X}$ and $\mathfrak{Y}$, it can be shown that this is a MARKOV Chain with transition probability matrix $\mathcal{P}$ and initial distribution λ, the same as $\mathfrak{X}$ (we shall provide an argument for this presently).

In particular, $\mathfrak{X}$ and $\mathfrak{Z}$ have the same finite-dimensional distributions.

• Using these remarks, we observe

$$\begin{aligned}
&\big|\,\mathbb{P}(X_n = j) - \pi_j\,\big| \;=\; \big|\,\mathbb{P}(Z_n = j) - \pi_j\,\big| \;=\; \big|\,\mathbb{P}(Z_n = j) - \mathbb{P}(Y_n = j)\,\big| \\
&= \big|\,\mathbb{P}(X_n = j\,,\, n < T) + \mathbb{P}(Y_n = j\,,\, n \geq T) - \mathbb{P}(Y_n = j)\,\big| \\
&= \big|\,\mathbb{P}(X_n = j\,,\, n < T) - \mathbb{P}(Y_n = j\,,\, n < T)\,\big| \;\leq\; 2\,\mathbb{P}(T > n) \;\longrightarrow\; 0\,, \quad \text{as } n \to \infty\,.
\end{aligned}$$

• To fill in the missing argument, observe that the strong MARKOV property applies to $\mathfrak{W} = \{W_0,\, W_1, \cdots\}$ with $W_n = (X_n, Y_n)$, and to the time T; this gives that $(X_{T+n}, Y_{T+n})_{n\in\mathbb{N}_0}$ is a MARKOV Chain with initial distribution $\delta_{(\mathfrak{b},\mathfrak{b})}$ and transition probability matrix $\widetilde{\mathcal{P}}$, and is independent of $(X_0, Y_0), \cdots, (X_T, Y_T)$. By symmetry, we can replace the sequence $(X_{T+n}, Y_{T+n})_{n\in\mathbb{N}_0}$ with $(Y_{T+n}, X_{T+n})_{n\in\mathbb{N}_0}$, which is also a MARKOV Chain with initial distribution $\delta_{(\mathfrak{b},\mathfrak{b})}$ and transition probability matrix $\widetilde{\mathcal{P}}$ and which remains independent of $(X_0, Y_0), \cdots, (X_T, Y_T)$.

Therefore, with the random sequence $\mathfrak{Z}' = \{Z'_0,\, Z'_1, \cdots\}$ defined via

$$Z'_n \;:=\; Y_n\,, \;\text{ for } n < T \qquad \text{and} \qquad Z'_n \;:=\; X_n\,, \;\text{ for } n \geq T\,,$$

the sequence $\mathfrak{W}' = \{W'_0,\, W'_1, \cdots\}$ with $W'_n = (Z_n, Z'_n)$, is a MARKOV Chain with state space $\mathcal{S} \times \mathcal{S}$, initial distribution μ, and transition probability matrix $\widetilde{\mathcal{P}}$.

In particular, it develops that $\mathfrak{Z} = \{Z_0,\, Z_1, \cdots\}$ is a MARKOV Chain with state space $\mathcal{S}$, initial distribution λ, and transition probability matrix $\mathcal{P}$. □

This impressive proof uses the method of "coupling" two independent MARKOV Chains. It is due to LINDVALL (1977) and is a huge improvement over earlier proofs of this result; those were analytic, and heavily dependent on the so-called "renewal theorem" (e.g., KARLIN & TAYLOR (1975)).

It goes completely off the tracks, if the chain is periodic; for instance, in the case of (12.15), if you start $\mathfrak{X} = \{X_0,\, X_1, \cdots\}$ with $X_0 = 1$, and $\mathfrak{Y} = \{Y_0,\, Y_1, \cdots\}$ with $\mathbb{P}(Y_0 = 1) = \mathbb{P}(Y_0 = 2) = 1/2$, the two chains will never meet on the event $\{Y_0 = 1\}$ and the argument just falls apart.

Proof of Theorem 12.12: Consider the "local" (or "total visitation") time $V_k(N) := \sum_{k=0}^{N-1} \mathbf{1}_{\{X_t = k\}}$ spent in state $k \in \mathcal{S}$ before day N. If k is transient, then we have almost surely

$$V_k \;:=\; \lim_{T\to\infty} V_k(N) \;<\; \infty\,, \qquad \text{thus} \qquad \lim_{N\to\infty} \frac{V_k(N)}{N} \;=\; 0 \;=\; \frac{1}{m_k}\,.$$

If, on the other hand, the state k is recurrent, we have $V_k = \infty$ and not only $\mathbb{P}(T_k < \infty) = 1$ but also $\mathbb{P}(T_k^{(r)} < \infty) = 1$, where $\{T_k^{(r)}\}_{r\in\mathbb{N}}$ is the sequence of successive visitation times of the chain to the state k in (12.7), and $\{S_k^{(r)}\}_{r\in\mathbb{N}}$ the corresponding sequence of excursion times in (12.8). These latter are independent random variables with the same distribution as T_k, thus in particular $\mathbb{E}(S_k^{(r)}) = m_k$.

Now let us note the almost sure comparisons

$$S_k^{(1)} + \cdots + S_k^{(V_k(N)-1)} \;\leq\; N - 1 < N \leq\; S_k^{(1)} + \cdots + S_k^{(V_k(N))}\,;$$

indeed, the leftmost term is the time of the last visit to k before day N, whereas the rightmost term is the time of the first visit to k after day $\mathbb{N}-1$. This gives

$$\frac{V_k(N)-1}{V_k(N)} \cdot \frac{S_k^{(1)}+\cdots+S_k^{(V_k(N)-1)}}{V_k(N)-1} \;\le\; \frac{N}{V_k(N)} \;\le\; \frac{S_k^{(1)}+\cdots+S_k^{(V_k(N))}}{V_k(N)}\,;$$

the claim (12.16) follows from this, the property $\mathbb{P}(V_k=\infty)=1$ and the Strong Law of Large Numbers

$$\lim_{N\to\infty} \frac{1}{N} \sum_{r=1}^{N} S_k^{(r)} \;=\; \frac{1}{m_k}\,.$$

Now property (12.16) leads to (12.17) for $f=\mathbf{1}_{\{k\}}$ for some state $k\in\mathcal{S}$. More generally, note that with $\pi_k=1/m_k$ we have that

$$\frac{1}{N}\sum_{t=0}^{N-1} f\big(X_t\big) - \sum_{k\in\mathcal{S}} \pi_k\, f(k) \;=\; \sum_{k\in\mathcal{S}} f(k)\,\left(\frac{V_k(N)}{N}-\pi_k\right) \tag{12.18}$$

holds almost surely, so the result follows immediately from (12.16) if the state space is finite, or at least if $\big|\,\{x\in\mathcal{S}\,:\,f(x)\neq 0\}\big|<\infty$. More generally, one tries to find a large finite set outside of which the sum on the right-hand side of (12.18) is small. The details are left as an exercise. □

Exercise 12.18. Provide the details of a rigorous argument, showing that the quantity in (12.18) tends to zero almost surely, as $n\to\infty$.

Under a few additional conditions, irreducible and positive recurrent MARKOV Chains obey suitable versions of the Central Limit Theorem. We simply state one such result right below, and refer to Chapter 1 of the book by KOMOROWSKI, LANDIM & OLLA (2012) for the proof.

Theorem 12.13. A CLT for MARKOV Chains: *Consider an irreducible and positive recurrent* MARKOV *Chain* $\mathfrak{X}=\{X_0,\,X_1,\cdots\}$, *possessing an invariant distribution* π *that satisfies the "detailed balance" conditions of Definition 12.6.*

Then for any $f:\mathcal{S}\to\mathbb{R}$ *with* $\sum_{i\in\mathcal{S}} f(i)\,\pi_i=0$ *and* $0<\sigma:=\big(\sum_{i\in\mathcal{S}} f^2(i)\,\pi_i\big)^{1/2}<\infty$, *the sequence*

$$\frac{1}{\sigma\sqrt{N}}\sum_{t=0}^{N-1} f\big(X_t\big)\,, \qquad N\in\mathbb{N}$$

converges in $\mathbb{P}_\pi-$*distribution to the standard Gaussian.*

13 Appendix: Elements of Combinatorial Analysis

The basic principle of Combinatorial Analysis is simple. Suppose you are given sets A_1, A_2, $\cdots$, A_n of finite cardinalities $|A_1| = m_1$, $|A_2| = m_2, \cdots$, $|A_n| = m_n$ and are asked to form ordered n-tuplets, or "vectors", of the form $(a_1, \cdots, a_n)$ with $a_1 \in A_1, \cdots, a_n \in A_n$. How many such vectors are there? Answer: $m_1 \cdot m_2 \cdots m_n$.

We present some very basic facts of Combinatorial Analysis, drawn from FELLER (1968) and ASH (1970).

13.1 Ordered Samples, with Replacement

In this spirit, suppose we are given n objects, and are asked to fill r boxes with items selected from the n objects, with replacement: (incarnations of) the same object can appear in multiple boxes. In how many ways can this be done?

Well, we can fill the first box in n ways, the second box in n ways, and so on to the last box. According to the principle above, we can make this selection in $\boxed{n^r}$ ways.

An *eleven-letter word,* for instance, is a sequence of $r = 11$ letters ("boxes") that have to be drawn ("filled") freely – that is, with replacement – from the Roman alphabet which contains $n = 26$ letters ("objects"). Order matters here, since ABRACADABRA is deemed rather different from ABARCADABAR. Clearly there are 26^{11} such words.

13.2 Ordered Samples, without Replacement: Permutations

Now suppose we are given again n objects, and are asked to fill r boxes ($2 \leq r \leq n$) with items selected from the n objects, but *without* replacement: once we have committed an object to a box, it gets locked there and cannot be placed again. In how many ways can this be done?

OK, now we can fill the first box in n ways, but the second box in only $n-1$ ways, the third box in $n-2$ ways, and so on to the last box. According to the basic principle, we can make this selection in

$$\boxed{(n)_r \,:=\, n(n-1)\cdots(n-r+1) \,=\, \frac{n!}{(n-r)!}}$$

ways. Each different such selection is called a *permutation* of r objects out of n. For instance, $26 \cdot 25 \cdots 15$ eleven-letter words in which no letter appears more than once, such as JOHANISBERG, can be constructed using the letters of the Roman alphabet.

With $r = n$ we get the cardinality

$$n\,! \,=\, n(n-1)\cdots 2 \cdot 1\,.$$

of the symmetric group Σ_n of all the permutations of n objects. We set, formally, $0! := 1$.

Exercise 13.1. The Birthday problem: A random sample *with* replacement is taken from a population of n elements, with all different possibilities equally likely. What is the probability that in the sample no element appears twice (to wit, that the sample could have been obtained also by sampling *without* replacement)?

For instance, in a class of r students what is the probability that there is no common birthday? Assuming that all possible arrangements are equally likely, the answer is

$$p \;=\; \frac{(n)_r}{n^r} \;=\; \frac{n(n-1)\cdots(n-r+1)}{n^r} \;=\; \left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\cdots\left(1-\frac{r-1}{n}\right)\,.$$

If we take $n = 365$ as the number of days in a year, and make the above assumption, then $p < 1/2$ for $r \geq 23$: *In a class of twenty three (or more) students, it is more likely than not, that at least two of them will have the same birthday.*

13.3 Unordered Samples, without replacement: Combinations

Now suppose you are in the situation of section 13.2, trying to fill r boxes with items taken from n objects, but you do not bother with the order in which the boxes get filled: if the objects are different kinds of ice cream flavors you select flavors to make yourself a sundae with r scoops ("boxes") *of different flavors* – which you then eat, and who cares about order. In how many ways can you make this sundae?

To put it a bit differently: how many *combinations* (unordered samples, without replacement) of size r can be formed using n objects? Or equivalently: *how many subsets of cardinality r does a given set of cardinality n possess, when $0 \leq r \leq n$?*

Suppose we single out one of these combinations; there are $r!$ ways to order its elements. If we do this exercise for each combination, we obtain all possible permutations (ways to choose ordered samples of size r) out of the n objects. Therefore, according to the basic principle once again, the total number of combinations that can be formed is given by the *Binomial coëfficient*

$$\boxed{\binom{n}{r} \;:=\; \frac{(n)_r}{r!} \;=\; \frac{n(n-1)\cdots(n-r+1)}{r!} \;=\; \frac{n!}{r!\,(n-r)!}\,.}$$

It is useful here to recall, for any given $n \in \mathbb{N}$, the *Binomial Theorem*

$$\big(a+b\big)^n \;=\; \sum_{k=0}^{n} \binom{n}{k}\, a^k\, b^{n-k}\,, \qquad\qquad (a,b) \in \mathbb{R}^2$$

and the PASCAL *Triangle Identity*

$$\binom{n-1}{r-1} + \binom{n-1}{r} \;=\; \binom{n}{r}\;, \quad r = 1, \cdots, n\,.$$

13.4 Occupancy Numbers

Suppose we have n distinct cells, and a collection of r balls which have to be placed in the cells. Each cell has the capacity to contain several balls, including all n of them. In how many ways can we allocate the balls into the cells?

If the balls are distinguishable, there are clearly n ways for each ball so, by the basic combinatorial principle, a total of n^r ways for them all.

In many cases, however, one has to treat the balls as *indistinguishable:* for instance, in statistical studies of accidents among weekdays, one is interested only in the number of occurrences and not in the identity of the individual involved in the accidents. Then all that matters is to know, for each cell $i = 1, \cdots, n$, its *occupancy number,* that is, the number $x_i \geq 0$ of balls it contains. Of course, $x_1 + \cdots + x_n = r$.

• Suppose that $r \geq n$ and that we insist that no cell can remain empty. *What then is the number of distinct vectors* $(x_1, \cdots, x_n) \in \mathbb{N}^n$ *with* $x_1 + \cdots + x_n = r$ *?* Let us represent balls by stars, and cells by the space between bars; for instance, let us consider the following configuration

$$| * *| * | * * * | * *| * |$$

with $r = 9$ balls and $n = 5$ boxes, thus $n + 1 = 6$ bars. Out of these bars, the first and last are fixed, so the remaining $n - 1$ bars have to be selected from among the $r - 1$ dividing points between different balls for the resulting configuration to "leave no box widowed". The number of distinct ways to do this, is given by

$$\binom{r-1}{n-1} = \frac{(r-1)!}{(n-1)!\,(r-n)!}\,.$$

• Now let us allow cells to remain empty; no assumption concerning the relative size of n and r is made. *What is the number of distinct vectors* $(x_1, \cdots, x_n) \in \mathbb{N}_0^n$ *with* $x_1 + \cdots + x_n = r$ *?* Again representing balls by stars and cells by the space between bars, let us consider the configuration

$$| * *| * | | | * *| * | | * |$$

with $r = 7$ balls and $n = 8$ boxes, thus $n + 1 = 9$ bars. Once again the first and last bars are set, but now the remaining $n - 1$ bars and the r balls can appear in an arbitrary order. In other words, we have to select, out of a total of $(n-1)+r$ symbols, the location of the r balls. This is sampling without regard to order and without replacement, so from section 13.3 the number is

$$\boxed{\binom{n-1+r}{r} = \frac{(n-1+r)!}{r!\,(n-1)!}\,.} \tag{13.1}$$

13.5 Unordered Samples, with replacement

How many samples of size r *can be constructed, if we are to pick the elements from a collection* $A = \{\alpha_1, \cdots, \alpha_n\}$ *of* n *items (alphabet)?* (No regard to order, no replacement, $n \geq r$.)

One way to approach this question is to count, for each element α_k of the set A, the number $x_k \in \{0, 1, \cdots, n\}$ of times it appears in the sample. Thus, the answer to this question is the total number of distinct vectors $(x_1, \cdots, x_n) \in \mathbb{N}_0^n$ with $x_1 + \cdots + x_n = r$. But this is precisely the question we answered in (13.1) above.

For instance, if we form eleven-letter collections using the $n = 26$ letters in the Roman alphabet but do not care about order (that is, we identify words such as ABRACADABRA and ABARCADABAR as one collection), there are $36 \cdot 35 \cdots 11$ such collections – as opposed to the 26^{11} words that can be formed when we insist on order (cf. section 13.1).

13.6 Multinomial Coëfficients

Suppose we play the game of *Astragaloi* of the ancients: [50] we cast a polyhedron with d facets $\mathfrak{f}_1, \cdots, \mathfrak{f}_d$ (say, a die if $k = 6$) a specified number n of times, and fix integers $r_1 \geq 0, \cdots, r_d \geq 0$ with $r_1 + \cdots + r_d = n$. *What is the number of outcomes that result in r_1 appearances of facet $\mathfrak{f}_1$, in r_2 appearances of facet $\mathfrak{f}_2$, and so on to r_k appearances of facet $\mathfrak{f}_d$?* (See Figure 3.)

Figure 3: Achilleus and Ajax throwing *Astragaloi. From the Exekias Amphora, 545-530 BC, the Vatican Museum.*

This is yet another application of the basic combinatorial principle. In order to effect the desired partition, we have to select the appearances of facet $\mathfrak{F}_1$, and this can be done in $(n)_{r_1}/r_1!$ ways; then we have to select the appearances of facet $\mathfrak{F}_2$, which can be done in $(n - r_1)_{r_2}/r_2!$ ways; and so on. According to the principle, the total number of favorable outcomes is

$$\frac{(n)_{r_1}}{r_1!} \cdot \frac{(n - r_1)_{r_2}}{r_2!} \cdot \frac{(n - r_1 - r_2)_{r_3}}{r_3!} \cdot \cdots \cdot \frac{(n - r_1 - r_2 - \cdots - r_{d-1})_{r_d}}{r_d!} = \frac{n!}{r_1!\, r_2! \cdots r_d!} .$$

For instance, at a Bridge table the 52 cards are partitioned into four equal groups, so the number of possible different configurations is $52!/(13!)^4 = (5.36) \times 10^{28}$. *What is the probability that each player gets an ace?* The four aces can be distributed in $4! = 24$ ways, while the remaining 48

[50] Games involving throwing "dice" were very popular in antiquity, but the Greeks did not roll our familiar, regular cubic dice. Instead, they played with *Astragaloi*, which had two rounded sides and only four playable surfaces, no two of which were identical. (This is perhaps a reason, why they failed to discern any regularity patterns in such throws.) Although the Greeks did believe that certain throws were more likely than others, these beliefs – superstitions we would call them nowadays – were not based on observation, and some were actually at variance with the actual likelihoods we could calculate today. Likewise, Roman emperors used lotteries to raise funds – but no attempt was made at the time to analyze them quantitatively. (Adapted from DEVLIN (2008), page 7.)

cards can be distributed in $48!/(12!)^4$ ways; thus, this probability is $24 \cdot 48! \cdot (13)^4/52! = 0.105$, close to 10%.

As another example, suppose now we throw twelve dice; *what is the probability that each face will appear twice?* There are 6^{12} possible outcomes, whereas the occurrence of the event under consideration can be cast as arranging 12 dice into six groups of two each; this can be done in $12!/(2! \cdots 2!) = 12!/2^6$ ways. Therefore, the probability we seek is $12!/(2^6 \cdot 6^{12}) = 0.003438$.

14 Appendix: The CARATHÉODORY-HAHN Construction

A systematic way of constructing measures according to certain "primitive" requirements is provided by the CARATHÉODORY-HAHN theory. This approach consists of

(i) the CARATHÉODORY Characterization Theorem 14.1, which identifies a σ-algebra, and a measure μ on it, from any *outer measure* μ^* (an outer measure is defined on all subsets of Ω, but is only countably *sub*-additive); of

(ii) the CARATHÉODORY Construction Theorem 14.2, which shows how an outer measure μ^* arises from any set-function $\nu : \mathcal{E} \to [0, \infty]$ defined on a family $\mathcal{E}$ of subsets of Ω with $\emptyset \in \mathcal{E}$, $\Omega \in \mathcal{E}$ and $\nu(\emptyset) = 0$; and of

(iii) the HAHN Restriction / Extension Theorem 14.3. This provides conditions under which μ^*, when restricted to the $\sigma-$algebra $\sigma(\mathcal{E})$ generated by the family $\mathcal{E}$, is an extension of ν; and shows that the extension is then essentially unique.

Taken together, these three results establish the validity of Theorem 6.1 which was stated in chapter 6 without proof. The LEBESGUE-STIELTJES measures outlined in the Examples of section 2.4 are obtained through this procedure, for instance by choosing $\mathcal{E}$ to be the algebra that consist of all finite, disjoint unions of half-open intervals $(a, b]$, and requiring $\nu((a, b\,]) = F(b) - F(a)$.

We fix throughout a given nonempty set Ω, and denote by $\mathcal{P}(\Omega)$ the collection of all its subsets. Our presentation will follow closely that of FOLLAND (1984).

Definition 14.1. Outer Measure: A set-function $\mu^* : \mathcal{P}(\Omega) \to [0, \infty]$ is said to be an *outer measure*, if

(i) $\mu^*(\emptyset) = 0$;

(ii) $\mu^*(E) \leq \mu^*(F)$ if $E \subseteq F$;

(iii) $\mu^*(\cup_{n\in\mathbb{N}} E_n) \leq \sum_{n\in\mathbb{N}} \mu^*(E_n)$ for any sequence $\{E_n\}_{n\in\mathbb{N}}$ of subsets of Ω.

(This is the so-called *countable sub-additivity property.*)

For instance, take $\Omega = \mathbb{N}$ and consider $\mu^*(A) = 0$ if $A = \emptyset$, $\mu^*(A) = 1$ if $A \neq \emptyset$, $A \subseteq \mathbb{N}$; this recipe defines an outer measure which is clearly not additive, even finitely. A systematic way of constructing outer measures, starting with set-functions that satisfy very minimal requirements, is outlined in Theorem 14.2 below.

Theorem 14.1. The CARATHÉODORY Characterization: *Let $\mu^* : \mathcal{P}(\Omega) \to [0, \infty]$ be an outer measure and consider the family $\mathcal{M}$ of subsets E of Ω which satisfy the condition*

$$\mu^*(A) \geq \mu^*(A \cap E) + \mu^*(A \cap E^c), \quad \forall\ A \subseteq \Omega\,. \tag{14.1}$$

Then $\mathcal{M}$ is a σ-algebra, and the restriction $\overline{\mu} = \mu^*\big|_{\mathcal{M}}$ of μ^* to $\mathcal{M}$ is a measure.

Clearly, the inequality (14.1) need be checked only for sets $A \subseteq \Omega$ with $\mu^*(A) < \infty$. Because of the sub-additivity property (iii) of Definition 14.1, the reverse inequality of (14.1) always holds; therefore, this condition is equivalent to

$$\mu^*(A) \;=\; \mu^*(A \cap E) + \mu^*(A \cap E^c), \quad \forall\ A \subseteq \Omega\,. \tag{14.2}$$

In words: every set $E \in \mathcal{M}$ has the property that it, and its complement, partition "cleanly" every subset A of Ω, without any loss of μ^*-mass.

Note also that for an outer-measure μ^* there can exist sets A and E with $\mu^*(A) < \mu^*(A \cap E) + \mu^*(A \cap E^c)$; for instance, in the example immediately following Definition 14.1, take $A = \{n-1, n, n+1\}$ and $E = \{n, n+1, \cdots\}$ for some $n \geq 2$.

Proof of Theorem 14.1: Evidently $\mathcal{M}$ contains the empty set and is closed under complementation, since (14.1) is symmetric in E, E^c. In fact, $\mathcal{M}$ contains every set $E \subseteq \Omega$ with $\mu^*(E) = 0$, as we then have $\mu^*(A \cap E) = 0$ for every $A \in \mathcal{P}(\Omega)$ and thus $\mu^*(A \cap E) + \mu^*(A \cap E^c) = \mu^*(A \cap E^c) \leq \mu^*(A)$, meaning $E \in \mathcal{M}$ on the strength of (14.1). (This observation points to the property of *completeness;* cf. Definition 14.2 and Exercise 14.2.)

We begin by showing that $\mathcal{M}$ is *closed under finite unions*. With E, F arbitrary elements of $\mathcal{M}$, and A any subset of Ω , we would like to show $E \cup F \in \mathcal{M}$. To this effect, we write $A \cap (E \cup F) = (A \cap E \cap F) \cup (A \cap E^c \cap F) \cup (A \cap E \cap F^c)$, so the subadditivity of μ^* implies

$$\mu^*(A \cap (E \cup F)) + \mu^*(A \cap (E \cup F)^c) \;\leq\; \mu^*(A \cap E \cap F) \tag{14.3}$$

$$+ \mu^*(A \cap E^c \cap F) + \mu^*(A \cap E \cap F^c) + \mu^*(A \cap E^c \cap F^c) \,.$$

But the condition (14.1) for E implies, respectively,

$$\mu^*(A \cap E \cap F) + \mu^*(A \cap E^c \cap F) \leq \mu^*(A \cap F) \,, \quad \mu^*(A \cap E \cap F^c) + \mu^*(A \cap E^c \cap F^c) \leq \mu^*(A \cap F^c) \,.$$

Substituting these in (14.3) and using (14.1) yet again, this time for F, we arrive at

$$\mu^*(A \cap (E \cup F)) + \mu^*(A \cap (E \cup F)^c) \leq \mu^*(A \cap F) + \mu^*(A \cap F^c) \leq \mu^*(A) \,.$$

We conclude that $E \cup F$ is in $\mathcal{M}$, and thus $\mathcal{M}$ is closed under pairwise unions, therefore also under finite unions.

• To show that $\mathcal{M}$ is also *closed under countable unions,* we consider next a sequence of sets $\{E_n\}_{n \in \mathbb{N}} \subseteq \mathcal{M}$. Their union can be written as the union of *pairwise disjoint* sets, namely

$$\bigcup_{n \in \mathbb{N}} E_n = \bigcup_{n \in \mathbb{N}} F_n \in \mathcal{M} \,, \quad \text{where} \quad F_n := E_n \setminus \Big(\bigcup_{k=1}^{n-1} E_k \Big) \tag{14.4}$$

for $n \geq 2$ and $F_1 := E_1$. Each set F_n is in $\mathcal{M}$, in view of the fact that $\mathcal{M}$ is closed under finite unions and under complementation.

Thus, we may assume that the $\{E_n\}_{n \in \mathbb{N}}$ are pairwise disjoint to begin with. It follows then from (14.2) that for any $A \subseteq \Omega$ with $\mu^*(A) < \infty$, we have

$$\begin{aligned} \mu^*(A \cap (\cup_{k=1}^n E_k)) &= \mu^*(A \cap (\cup_{k=1}^n E_k) \cap E_n) + \mu^*(A \cap (\cup_{k=1}^n E_k) \cap E_n^c) \\ &= \mu^*(A \cap E_n) + \mu^*(A \cap (\cup_{k=1}^{n-1} E_k)), \qquad \forall\, n \in \mathbb{N} \,. \end{aligned}$$

This can be written equivalently as $\mu^*(A \cap (\cup_{k=1}^n E_k)) - \mu^*(A \cap (\cup_{k=1}^{n-1} E_k)) = \mu^*(A \cap E_n)$ (the assumption $\mu^*(A) < \infty$ is crucial here), and gives

$$\mu^*(A \cap (\cup_{k=1}^n E_k)) \;=\; \sum_{k=1}^n \mu^*(A \cap E_k), \quad \forall\, n \in \mathbb{N}\,. \tag{14.5}$$

Now (14.2), (14.5) and the monotonicity of μ^* imply

$$\mu^*(A) \;=\; \mu^*\big(A \cap (\cup_{k=1}^n E_k)\big) + \mu^*\big(A \cap (\cup_{k=1}^n E_k)^c\big) \geq \sum_{k=1}^n \mu^*(A \cap E_k) + \mu^*(A \cap (\cup_{k\in\mathbb{N}} E_k)^c)\,. \tag{14.6}$$

Letting as $n \to \infty$ in (14.6), we obtain from the countable subadditivity of μ^* that

$$\begin{aligned} \mu^*(A) &\geq \sum_{k\in\mathbb{N}} \mu^*(A \cap E_k) + \mu^*\Big(A \cap \Big(\bigcup_{k\in\mathbb{N}} E_k\Big)^c\Big) \\ &\geq \mu^*\Big(A \cap \Big(\bigcup_{k\in\mathbb{N}} E_k\Big)\Big) + \mu^*\Big(A \cap \Big(\bigcup_{k\in\mathbb{N}} E_k\Big)^c\Big)\,. \end{aligned} \tag{14.7}$$

We have proved (14.7) — that is, (14.1), or equivalently (14.2), for $E = \bigcup_{k\in\mathbb{N}} E_k$ — assuming $\mu^*(A) < \infty$; but of course the inequality (14.7) holds also if $\mu^*(A) = \infty$, thus for any $A \subseteq \Omega$. This shows that $\mathcal{M}$ is a σ-algebra.

Taking now $A = \bigcup_{k\in\mathbb{N}} E_k$ in the first inequality of (14.7) — actually written as equality, on accout of the equivalence between (14.1) and (14.2) — leads to the identity

$$\mu^*\Big(\bigcup_{k\in\mathbb{N}} E_k\Big) \;=\; \sum_{k\in\mathbb{N}} \mu^*(E_k)\,,$$

so that μ^* is countably additive on $\mathcal{M}$. In other words, the restriction $\overline{\mu} = \mu^*\big|_{\mathcal{M}}$ is a measure. □

The next result shows how to construct outer measures from set functions that satisfy *very minimal requirements,* through a procedure of what can be thought of as "least expensive coverage". This construction justifies also the terminology *outer measure*.

Theorem 14.2. The CARATHÉODORY Construction of Outer Measures: *Let $\mathcal{E}$ be a family of subsets of Ω, which includes Ω as well as the empty set $\emptyset$.*

Given any set-function $\nu : \mathcal{E} \to [0,\infty]$ satisfying $\nu(\emptyset) = 0$, the set-function

$$\mu^*(A) \;:=\; \inf\left\{ \sum_{n\in\mathbb{N}} \nu(E_n) \;\middle|\; \{E_n\}_{n\in\mathbb{N}} \subseteq \mathcal{E}\,,\; A \subseteq \bigcup_{n\in\mathbb{N}} E_n \right\}, \qquad A \subseteq \Omega \tag{14.8}$$

defines an outer measure on $\mathcal{P}(\Omega)$, called the "outer measure generated by ν".

Proof: By taking $E_n \equiv \emptyset$ in (14.8), we see that $\mu^*(\emptyset) = 0$. Also $\mu^*(A) \leq \mu^*(B)$ whenever $A \subseteq B$, since any covering of B is also a covering of A.

Finally, let $\{A_k\}_{k\in\mathbb{N}}$ be a sequence of subsets of Ω, take any $\varepsilon > 0$, and recall the definition of the infimum, or "greatest lower bound", of a set. For each $k \in \mathbb{N}$, select a covering $\{E_{k,n}\}_{n\in\mathbb{N}} \subseteq \mathcal{E}$

of A_k, that is, $A_k \subseteq \bigcup_{n\in\mathbb{N}} E_{k,n}$, such that $\sum_{n\in\mathbb{N}} \nu(E_{k,n}) \leq \mu^*(A_k) + \varepsilon 2^{-k}$. Then $\bigcup_{k\in\mathbb{N}} A_k \subseteq \bigcup_{k\in\mathbb{N}} \bigcup_{n\in\mathbb{N}} E_{k,n}$ and

$$\mu^*\left(\bigcup_{k\in\mathbb{N}} A_k\right) \leq \sum_{k\in\mathbb{N}} \sum_{n\in\mathbb{N}} \nu(E_{k,n}) \leq \sum_{k\in\mathbb{N}} \mu^*(A_k) + \varepsilon.$$

Since ε was arbitrary, we conclude that μ^* is countably subadditive. □

Theorem 14.2 associates an outer measure μ^* to *any* given set-function ν that satisfies its conditions. But the outer measure μ^* may be quite different from ν, even when restricted to sets in the original family $\mathcal{E}$. The following theorem gives conditions, under which ν can be extended to a measure on $\sigma(\mathcal{E})$, which agrees with ν when restricted back to $\mathcal{E}$; and that this extension can be done in a unique manner.

Theorem 14.3. HAHN Restriction / Extension: *Suppose that in Theorem 14.2 the family $\mathcal{E}$ is an algebra, and the set-function $\nu : \mathcal{E} \to [0,\infty]$ a* pre-measure *on $\mathcal{E}$. Denote by μ^* the outer measure generated by ν as in (14.8), and by*

$$\mu := \mu^*\big|_{\sigma(\mathcal{E})}$$

the restriction of this outer measure to the σ-algebra $\sigma(\mathcal{E})$ generated by $\mathcal{E}$ *(cf. Theorem 6.1). Let us recall also the σ-algebra $\mathcal{M}$ of Theorem 14.2 associated with the outer measure μ^*.*

(i) *Then $\sigma(\mathcal{E}) \subseteq \mathcal{M}$, and $\mu\big|_{\mathcal{E}} = \mu^*\big|_{\mathcal{E}} = \nu$.*

(ii) *For any measure ρ on $\sigma(\mathcal{E})$ which satisfies $\rho\big|_{\mathcal{E}} = \nu$, we have*

$$\rho(A) \leq \mu(A), \quad \forall\, A \in \sigma(\mathcal{E}) \qquad \textit{(in fact with equality, with equality when } \mu(A) < \infty).$$

(iii) *If ν is* σ-finite, *then μ is the unique measure on $\sigma(\mathcal{E})$ with $\mu\big|_{\mathcal{E}} = \nu$; to wit, the unique extension of the pre-measure ν on $\mathcal{E}$ to a measure on $\sigma(\mathcal{E})$.*

Proof: We start with *Part (i)*. Since $\mathcal{M}$ is a σ-algebra, to prove $\sigma(\mathcal{E}) \subseteq \mathcal{M}$ it is enough to show $\mathcal{E} \subseteq \mathcal{M}$, i.e., that every set $E \in \mathcal{E}$ satisfies (14.1) for any $A \subseteq \Omega$. To this end, fix $\varepsilon > 0$ and find a sequence $\{Z_k\}_{k\in\mathbb{N}} \subseteq \mathcal{E}$ with $A \subseteq \cup_{k\in\mathbb{N}} Z_k$ and

$$\mu^*(A) + \varepsilon \geq \sum_{k\in\mathbb{N}} \nu(Z_k) = \sum_{k\in\mathbb{N}} [\nu(Z_k \cap E) + \nu(Z_k \cap E^c)] \geq \mu^*(A \cap E) + \mu^*(A \cap E^c).$$

Here we made use of the definition of outer measure, of the fact that $Z_k \cap E$ and $Z_k \cap E^c$ belong to the algebra $\mathcal{E}$, and of the fact that ν is a pre-measure on $\mathcal{E}$. Since $\varepsilon > 0$ was arbitrary, it follows that $E \in \mathcal{M}$.

- We show next that

$$\mu^*(A) = \nu(A)$$

holds for every $A \in \mathcal{E}$. The inequality $\mu^*(A) \leq \nu(A)$ is obvious, since we can construct a covering of A in (3.7) using sets in $\mathcal{E}$, simply by taking $E_1 = A$, $E_n = \emptyset$ for $n \geq 2$.

To argue the reverse inequality, let $\{E_k\}_{k\in\mathbb{N}} \subseteq \mathcal{E}$ be any covering of A, namely $A \subseteq \cup_{k\in\mathbb{N}} E_k$, with sets in $\mathcal{E}$. Then

$$F_n := A \cap E_n \cap \Big(\bigcup_{k=1}^{n-1} E_k\Big)^c, \quad n \in \mathbb{N}$$

are pairwise disjoint sets in $\mathcal{E}$, whose union $\bigcup_{n\in\mathbb{N}} F_n = A$ belongs to the algebra $\mathcal{E}$. Now, the fact that ν is a pre-measure on $\mathcal{E}$ gives

$$\nu(A) \,=\, \nu\Big(\bigcup_{n\in\mathbb{N}} F_n\Big) \,=\, \sum_{n\in\mathbb{N}} \nu(F_n) \,\leq\, \sum_{n\in\mathbb{N}} \nu(E_n)\,.$$

From this inequality, and the arbitrariness of the sequence $\{E_k\}_{k\in\mathbb{N}} \subseteq \mathcal{E}$ that covers the set A, we obtain $\nu(A) \leq \mu^*(A)$. Therefore $\nu(A) = \mu^*(A)$, so *Part (i)* is proved.

- To establish *Part (ii),* let $A \in \sigma(\mathcal{E})$ and suppose that

$$\{E_k\}_{k\in\mathbb{N}} \subseteq \mathcal{E} \quad \text{is an } \textit{arbitrary} \text{ sequence with } A \subseteq \bigcup_{k\in\mathbb{N}} E_k =: E\,. \tag{14.9}$$

Then we have

$$\rho(A) \leq \rho(E) \leq \sum_{k\in\mathbb{N}} \rho(E_k) = \sum_{k\in\mathbb{N}} \nu(E_k)$$

from the countable subadditivity of ρ, and the fact that ρ and ν agree on $\mathcal{E}$. Because the sequence $\{E_k\}_{k\in\mathbb{N}} \subseteq \mathcal{E}$ is arbitrary, we deduce $\rho(A) \leq \mu^*(A) = \mu(A)$ from (14.8) and $A \in \sigma(\mathcal{E})$, $\mu := \mu^*\big|_{\sigma(\mathcal{E})}$. This shows that μ is the "maximal extension" of ν to a measure on $\sigma(\mathcal{E})$.

- Let us refine this result when $A \in \sigma(\mathcal{E})$ satisfies $\mu(A) = \mu^*(A) < \infty$. Then, given any $\varepsilon > 0$, we can choose the sets in the covering sequence $\{E_k\}_{k\in\mathbb{N}} \subseteq \mathcal{E}$ of (14.9) to be pairwise disjoint and to satisfy the inequality $\sum_{k\in\mathbb{N}} \nu(E_k) \leq \mu^*(A) + \varepsilon$ (recall again the definition of the infimum of a set). This, in turn, leads to the string of inequalities

$$\mu(A) \leq \mu(E) \leq \sum_{k\in\mathbb{N}} \mu(E_k) = \sum_{k\in\mathbb{N}} \rho(E_k) = \sum_{k\in\mathbb{N}} \nu(E_k) \leq \mu^*(A) + \varepsilon = \mu(A) + \varepsilon\,;$$

in particular, since $\mu(E) = \mu(A) + \mu(E \setminus A)$ and $\mu(A) < \infty$, we get from this $\mu(E \setminus A) \leq \varepsilon$ by comparing the second and last terms.

We recall now that both sets A, E (thus also $E \setminus A$) are in $\sigma(\mathcal{E})$. On the strength of the countable additivity of ρ, and of the dominance $\rho \leq \mu$ we already proved, we deduce

$$\mu(A) \leq \mu(E) \leq \sum_{k\in\mathbb{N}} \rho(E_k) = \rho(E) = \rho(A) + \rho(E \setminus A) \leq \rho(A) + \mu(E \setminus A) \leq \rho(A) + \varepsilon.$$

Since this is true for any $\varepsilon > 0$, it gives $\mu(A) \leq \rho(A)$. We have already established the reverse inequality, so $\mu(A) = \rho(A)$.

- *Part (iii),* the last statement of the theorem, follows now easily: in the case of σ-finite ν we have $\Omega = \cup_{k\in\mathbb{N}}\Omega_k$, where the sets $\{\Omega_k\}_{k\in\mathbb{N}} \subseteq \mathcal{E}$ are pairwise disjoint with $\mu(\Omega_k) = \nu(\Omega_k) < \infty$ for every $k \in \mathbf{N}$, and thus $\mu(A) = \sum_{k\in\mathbb{N}} \mu(A \cap \Omega_k) = \sum_{k\in\mathbb{N}} \rho(A \cap \Omega_k) = \rho(A)$ for any $A \in \sigma(\mathcal{E})$. The proof of the Theorem is complete. □

14.1 Completeness of Measure Spaces

Occasionally we want to show that some given function g is $\mathcal{F}-$measurable. We identify a function f which we know has this property, and then manage to show that $\{f \neq g\}$ is (contained in) a set of zero measure.

Can we conclude that g is $\mathcal{F}-$measurable from this? The answer is *yes, if the measure space has the property of* **completeness**; see Definition 14.2 below. We introduce this very important notion in this section, and study its ramifications in a series of exercises.

Definition 14.2. Null Sets, Negligible Sets, Completeness: Consider a measure space $(\Omega, \mathcal{F}, \mu)$. A set $E \in \mathcal{F}$ is called *null* for the measure μ, if $\mu(E) = 0$.

We say that a subset $F \subseteq \Omega$ is *μ-negligible*, and write $F \in \mathcal{N}$, if there exists a null set E with $F \subseteq E$. A given statement is said to hold *$\mu-$almost everywhere (a.e.)*, if the set on which it fails is negligible.

Two classes $\mathcal{A}, \mathcal{B}$ of subsets of Ω are said to *agree modulo* μ, and we write $\mathcal{A} = \mathcal{B}$ modulo μ, if $\mathcal{A} \setminus \mathcal{B} \subseteq \mathcal{N}$ and $\mathcal{B} \setminus \mathcal{A} \subseteq \mathcal{N}$.

Finally, we say that the measure(-space) is *complete*, if $\mathcal{N} \subseteq \mathcal{F}$. □

By the subadditivity of μ, any countable union of null sets is again null. The union of an uncountable collection of null sets can easily fail to be a null set.

Completeness is a very useful property for a measure space (see Exercise 14.3, for examples) and simplifies many technical arguments. It can always be achieved by suitably enlarging the domain of μ, as the following exercises demonstrate.

Exercise 14.1. Completion of a Measure Space: (i) In the context of Definition 14.2, show that $\overline{\mathcal{F}} := \{E \cup F \,|\, E \in \mathcal{F},\ F \in \mathcal{N}\}$ is a σ-algebra, that $\overline{\mu}(E \cup F) := \mu(E)$ defines well a complete measure on $\overline{\mathcal{F}}$, and that this $\overline{\mu}$ is the *unique* extension of μ to a complete measure on $\overline{\mathcal{F}}$.

The measure space $(\Omega, \overline{\mathcal{F}}, \overline{\mu})$ is called the *completion of* $(\Omega, \mathcal{F}, \mu)$; these two spaces coincide, if the original measure-space is complete.

(ii) If $f : \Omega \to \mathbb{R}$ is $\overline{\mathcal{F}}-$measurable, then there exists an $\mathcal{F}-$measurable function $g : \Omega \to \mathbb{R}$ such that $f = g$, $\overline{\mu}-$a.e.

Exercise 14.2. Completeness in the CARATHÉODORY construction:

(i) Show that the measure $\overline{\mu} := \mu^*\big|_{\mathcal{M}}$ of Theorem 14.1 is complete.

(ii) If μ^* is the outer-measure generated by a pre-measure ν on an algebra $\mathcal{E}$, in the manner of Theorems 14.2 and 14.3, then this measure's restriction $\mu^*\big|_{\sigma(\mathcal{E})} =: \mu$ to $\sigma(\mathcal{E})$ need not be complete.

(iii) Show, however, that the completion of the measure space $(\Omega, \sigma(\mathcal{E}), \mu)$ is the measure space $(\Omega, \mathcal{M}, \overline{\mu})$ where $\overline{\mu} := \mu^*\big|_{\mathcal{M}}$ and μ^* is the outer measure of part (i), provided this outer measure is σ-finite.

(*Hint:* Every $A \subseteq \Omega$ has a "measurable cover", that is, a set $B \in \mathcal{F}$ with $A \subseteq B$, $\mu^*(A) = \mu^*(B)$. And every $E \in \mathcal{M}$ can be written as $E = B \cup N$ with $B \in \mathcal{F}$, $N \subseteq D$ for some $D \in \mathcal{F}$ with $\mu^*(D) = 0$.)

Exercise 14.3. The benefits of Completeness: Let $(\Omega, \mathcal{F}, \mu)$ be a *complete* measure space, and consider real-valued functions $f,\ g,\ \{f_n\}_{n\in\mathbb{N}}$ on it.

(i) If f is measurable and $f = g$ holds $\mu-$a.e., then g is also measurable.

(ii) If $\{f_n\}_{n\in\mathbb{N}}$ are measurable and $\lim_n f_n = f$ holds $\mu-$a.e., then f is also measurable.

14.2 LEBESGUE Measure

According to Exercise 14.2, the completion of the measure space $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mu_F)$ constructed in chapter 6 is the σ-finite measure space $(\mathbb{R}, \mathcal{M}_F, \overline{\mu}_F)$. The measure $\overline{\mu}_F$ is regular, in the sense of Exercise 4.26: it is called the **LEBESGUE-STIELTJES measure** induced by the distribution function F.

This measure is defined on the σ-algebra $\mathcal{M}_F \equiv \mathcal{M}$ of so-called $F-$LEBESGUE-STIELTJES *measurable sets,* which is constructed as in Theorems 14.1–14.3 and satisfies

$$\mathcal{B}(\mathbb{R}) \subset \mathcal{M}_F \subset \mathcal{P}(\mathbb{R})\,. \tag{14.10}$$

Here $\mathcal{P}(\mathbb{R})$ is the collection of all subsets of $\mathbb{R}$. We shall see in chapter 15 (Appendix) that both inclusions in (14.10) are typically strict.

For the choice $F(x) \equiv x$, the resulting $\overline{\lambda} \equiv \overline{\mu}_F$ is the **LEBESGUE measure on the real line**, and the class $\mathcal{L} \equiv \mathcal{M}_F$ is the *σ-algebra of* LEBESGUE-*measurable sets.* This class is invariant under translations and dilations, and so is LEBESGUE measure, in the sense that $E + s \in \mathcal{L}$, $rE \in \mathcal{L}$ and

$$\overline{\lambda}(E + s) = \overline{\lambda}(E)\,, \quad \overline{\lambda}(r\,E) = r\,\overline{\lambda}(E)$$

for every $E \in \mathcal{L}\,,\ s \in \mathbb{R}\,,\ r > 0\,.$

The restriction $\lambda = \overline{\lambda}\big|_{\mathcal{B}(\mathbb{R})}$ to the σ-algebra of BOREL sets, is also called "LEBESGUE measure"; it has the property $\lambda((a, b]) = b - a\,.$

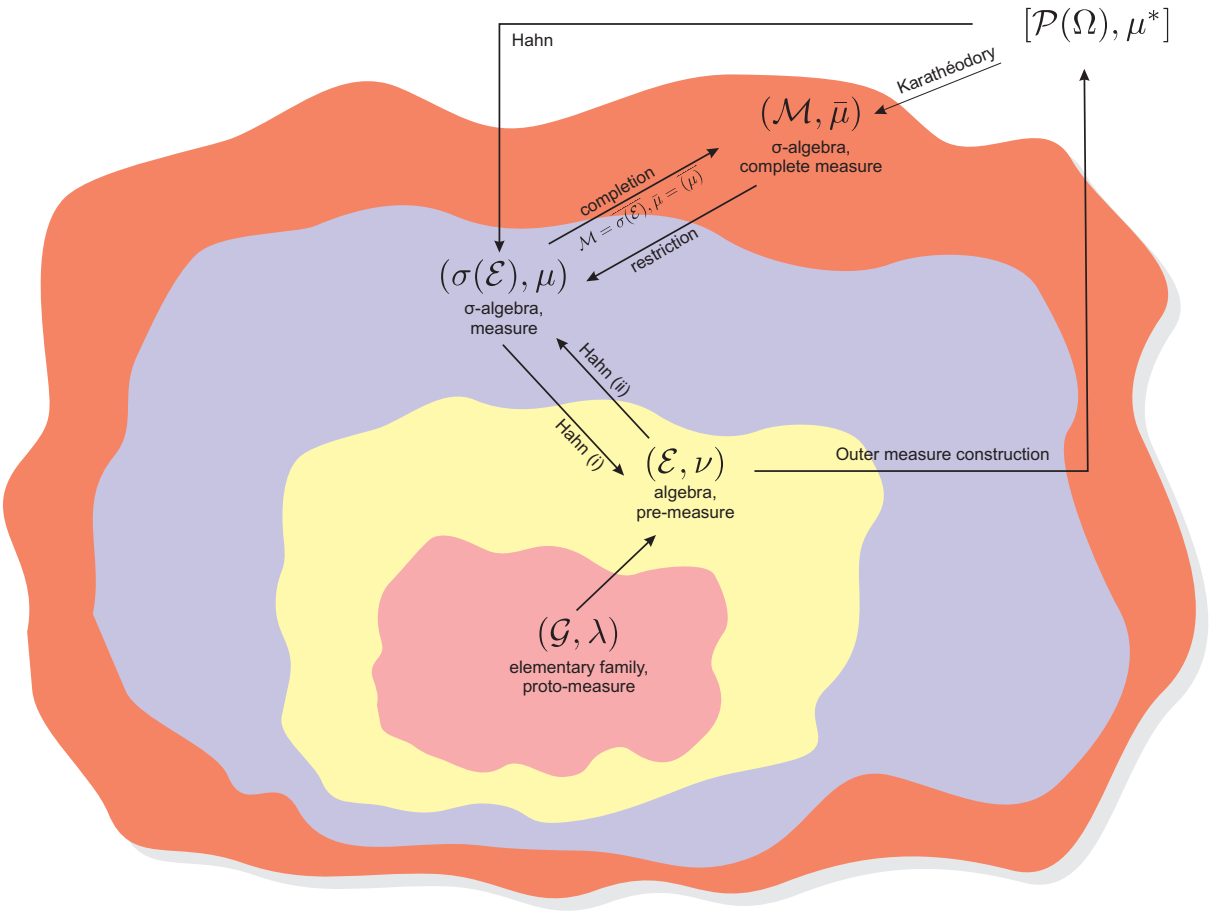

Figure 4: From proto-measures on elementary families, to pre-measures on algebras, to outer measures on power sets, down to measures on sigma-algebras and then up to complete measures.

15 Appendix: The CANTOR Set and Function

We construct in this section the celebrated CANTOR *set* and its associated CANTOR *function.* These provide a rich source for the understanding of LEBESGUE measure, and of notions such as singularity and absolute continuity.

Let us start by considering the unit interval $C_0 = [0,1]$, from which we then remove the middle third open interval to obtain $C_1 = [0,1/3] \cup [2/3,1]$. We do the same to each of the two closed intervals that comprise C_1, and obtain $C_2 = [0,1/9] \cup [2/9,1/3] \cup [2/3,7/9] \cup [8/9,1]$. Continuing this process inductively, we see that C_n will contain 2^n such closed intervals; by removing the middle third from each of these, we obtain C_{n+1}.

The CANTOR set is the countable intersection $C := \bigcap_{n\in\mathbb{N}} C_n$.

This set is clearly *non-empty* and *compact,* as it is the intersection of a decreasing sequence of compact intervals. In particular it is a closed, therefore BOREL, set, and is represented as

$$C = \left\{ x \in [0,1] \;\middle|\; x = \sum_{k\in\mathbb{N}} \frac{a_k}{3^k}\,, \quad \text{where } a_k = 0 \text{ or } 2\,, \;\; \forall\; k \in \mathbb{N} \right\}. \tag{15.1}$$

Indeed, the left-endpoints of the 2^n intervals comprising C_n are the points of the form $\sum_{k=1}^{n} a_k\, 3^{-k}$, where $a_k \in \{0,2\}$. This can be seen easily, by induction. Thus, a number of the form $x = \sum_{k\in\mathbb{N}} a_k\, 3^{-k}$, where $a_k \in \{0,2\}$, is in all the sets C_n, therefore in C; and any point in the set

$$G \;:=\; [0,1] \setminus C \;=\; \bigcup_{n\in\mathbb{N}} \bigcup_{m=1}^{2^{n-1}} G_n^m \tag{15.2}$$

cannot have such a ternary expansion.

We have denoted here by $\{G_n^m \,|\, 1 \le m \le 2^{n-1}\}_{n\in\mathbb{N}}$ the collection of disjoint open intervals that have to be removed from $[0,1]$ to obtain C, namely

$$G_1^1 = (1/3, 2/3)$$

$$G_2^1 = (1/9, 2/9)\,, \qquad G_2^2 = (7/9, 8/9)\,,$$

$$G_3^1 = (1/27, 2/27)\,,\;\; G_3^2 = (7/27, 8/27)\,,\;\; G_3^3 = (19/27, 20/27)\,,\;\; G_3^4 = (25/27, 26/27)\,, \cdots$$

et cetera. Furthermore, we observe that

$$C \text{ is } \textit{"dense in itself":} \text{ it coincides with the set } C' \text{ of its cluster points.} \tag{15.3}$$

To see this, take any $x = \sum_{k\in\mathbb{N}} a_k\, 3^{-k} \in C$ and, for each $k \in \mathbb{N}$ separately, change a_k from 0 to 2 or vice-versa. We obtain a sequence $\{x_k\}_{k\in\mathbb{N}}$ of distinct points in C, that converges to x : $C \subseteq C'$. Closedeness gives $C' \subseteq C$, and the conclusion follows. Note, however, that *between any two points in* $x,\, y$ *in* C, *lies an open interval contained in* $G = [0,1] \setminus C$; indeed, if $|x - y| > 3^{-n}$, then these two points belong to different intervals C_n of C. In addition to all this,

$$\textit{the LEBESGUE measure of the CANTOR set is equal to zero:}\;\; \lambda(C) = 0\,, \tag{15.4}$$

since $1 - \lambda(C) = \sum_{n\in\mathbb{N}} \sum_{m=1}^{2^{n-1}} \lambda(G_n^m) = \sum_{n\in\mathbb{N}} 2^{n-1} \cdot 3^{-n} = \frac{1}{3} + \frac{2}{9} + \frac{4}{27} + \cdots = 1$.

15.1 The Devil's Staircase

Consider now the function $F : C \to [0,1]$ defined by

$$F(x) = \sum_{k\in\mathbb{N}} \frac{b_k}{2^k} \quad \text{with } b_k = \frac{a_k}{2} \in \{0,1\}, \quad \text{for } x = \sum_{k\in\mathbb{N}} \frac{a_k}{3^k} \in C \text{ and } a_k \in \{0,2\}. \tag{15.5}$$

To begin to get feel for this definition, note that $x = 0$ is of the form (15.1) with $a_k = 0$ for all $k \in \mathbb{N}$; the number $x = 1/3$ is of this form with $a_1 = 0$ and $a_k = 2$ for all $k \geq 2$; the number $x = 2/3$ is again of the form (15.1) with $a_1 = 2$ and $a_k = 0$ for all $k \geq 2$; and $x = 1$ is of this form with $a_k = 2$ for all $k \in \mathbb{N}$. For these numbers, the values of the function $F(\cdot)$ are given as

$$F(0) = 0\,, \quad F(1/3) = \sum_{k\geq 2} \frac{1}{2^k} = \frac{1}{2} = F(2/3)\,, \quad F(1) = \sum_{k\in\mathbb{N}} \frac{1}{2^k} = 1\,.$$

Note that the expression for $F(x)$ in (15.5) is the binary expansion of some number in the interval $[0,1]$; whereas any $z \in [0,1]$ can be obtained this way (as $z = F(x)$ for some $x \in C$). Therefore,

$$C \text{ can be mapped } \textit{onto } [0,1]\,, \text{ and is thus uncountable.} \tag{15.6}$$

Setting $F(x) := (2m-1)/2^n$ for every $x \in G_n^m$, we extend the function $F(\cdot)$ of (15.5) to all of $[0,1]$. You can visualize this by imagining that each of the intervals G_n^m, $m = 1, \cdots, 2^{n-1}$ that you delete at stage $n \in \mathbb{N}$ in order to form the CANTOR set, is not thrown away but is instead "hoisted" at height $(2m-1)/2^n$: the middle interval $(1/3, 2/3)$ at height $1/2$; the interval $(1/9, 2/9)$ at height $1/4$, whereas the interval $(7/9, 8/9)$ at height $3/4$; and so on.

The resulting, so-called **CANTOR function**

- is *increasing* (because $x < y$ implies $F(x) < F(y)$, unless x, y are endpoints of one of the open intervals $\{G_n^m\,;\, m = 1, \cdots, 2^{n-1}\}_{n\in\mathbb{N}}$ in (15.2), in which case $F(x) = F(y)$);
- satisfies $F(0) = 0,\ F(1) = 1$;
- is *continuous* (its range is $[0,1]$, so it cannot have jumps); and
- is *flat on* $G = [0,1] \setminus C$ (by construction).

Therefore, the derivative of $F(\cdot)$ exists and is equal to zero everywhere on the set G, which has full Lebesgue measure; thus *it is impossible to write* $F(\cdot)$ *in the form* $F(x) = \int_0^x f(u)\,du$ for some Borel-measurable $f : [0,1] \to [0,\infty)$ (because this would imply $f = 0,\ \lambda-$a.e. on $[0,1]$, and thus $F(1) = 0$, violating $F(1) = 1$).

In other words: the **CANTOR measure** $\kappa = \mu_F$ induced by the CANTOR function $F(\cdot)$ on $\mathcal{B}([0,1])$, is *singular with respect to* LEBESGUE *measure.*

Quite a few writers refer to the graph of the CANTOR function as the "Devil's staircase". This structure is clearly odd: it looks full, totally opaque, if you stare at it looking upward (because then you are facing the intervals that constitute the set G of (15.2), which has measure one); but it looks empty, or diaphanous at most, when you stare at it sideways (because then you are facing the dust-like collection of its "stairs" os "slugs", the places where the case goes up, and these are located on the CANTOR set C which has measure zero). As MANDELBROT (1982), page 82,

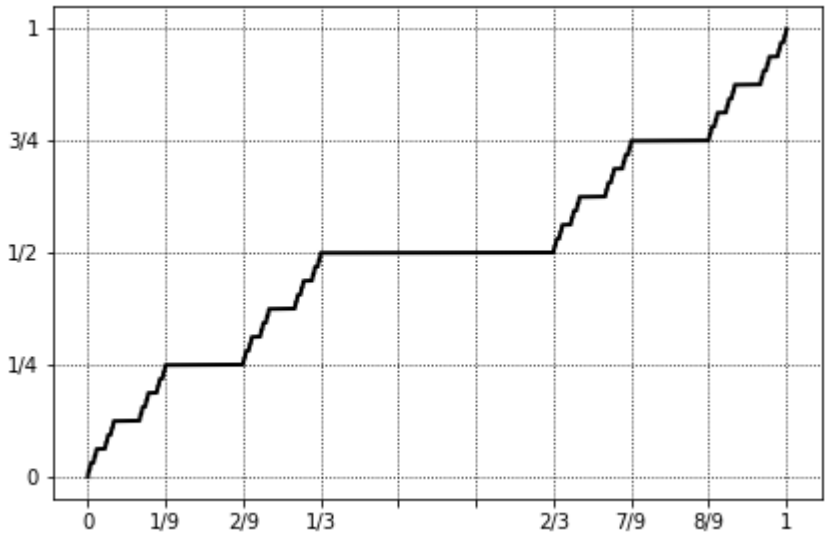

Figure 5: Graph of the CANTOR function.

puts it: "(this function) must manage to increase *somewhere* from the point of coördinates (0,0) to the point of coördinates (1,1). It increases over infinitely many, infinitely small, highly clustered jumps, corresponding to the slugs". Read the discussion on pages 79-84 of this book, and marvel at the pictures it provides.

15.2 Non-Measurable Sets

With the help of these notions, we can show now that there exist subsets of the real line that are not BOREL measurable. The arguments, however, will be highly non-constructive, and will rely on the *Axiom of Choice.* [51]

Proposition 15.1. *There exist subsets of the real line that are not* LEBESGUE*-measurable.*

Proof : Consider the interval $[0,1)$ and define the equivalence relation $x \sim y \;\Leftrightarrow\; x - y \in \mathbf{Q}$, where $\mathbf{Q}$ is the set of rational numbers and $\mathbf{Q}_1 := \mathbf{Q} \cap [0,1)$. From the Axiom of Choice, there exists a set $E \subset [0,1)$ that contains precisely one member from each of the resulting equivalence classes. For each $q \in \mathbf{Q}_1$, "move E to the right by q units, and then move back to the beginning of the interval the part that sticks out", to create the new set

$$E_q := \{y + q \,|\, y \in E\,,\; 0 \le y < 1 - q\} \cup \{y + q - 1 \,|\, y \in E\,,\; 1 - q \le y < 1\}\,. \tag{15.7}$$

Then every $x \in [0,1)$ belongs to exactly one of the E_q's (exercise).

We claim that *the set E does not belong to the σ-algebra $\mathcal{L}$ of* LEBESGUE*-measurable sets.* Suppose the contrary; then $\{E_q\}_{q \in \mathbf{Q}_1} \subseteq \mathcal{L}$ (because $\mathcal{L}$ is invariant under translations) and consequently $[0,1) = \bigcup_{q \in \mathbf{Q}_1} E_q \in \mathcal{L}$, whence the absurdity

$$1 = \bar{\lambda}([0,1)) = \sum_{q \in \mathbf{Q}_1} \bar{\lambda}(E_q) = \sum_{q \in \mathbf{Q}_1} \bar{\lambda}(E)\,. \tag{15.8}$$

Here the last summation is either zero (if $\bar{\lambda}(E) = 0$) or infinite (if $\bar{\lambda}(E) > 0$). □

It was shown by SOLOVAY (1970) that it is just not possible to construct non-LEBESGUE-measurable sets without invoking the Axiom of Choice.

Proposition 15.2. *There exist* LEBESGUE*-measurable subsets of the real line that are not* BOREL *sets.*

Proof : Consider the continuous, strictly increasing function $h(x) := x + F(x),\; 0 \le x \le 1$, and note that $h(B) \in \mathcal{B}(\mathbb{R})$ for every BOREL set $B \in \mathcal{B}(\mathbb{R})$. For each of the intervals G_n^m of (15.2), its image $h(G_n^m)$ is an interval of the same length (because $F(\cdot)$ is flat on G), so that $\overline{\lambda}(h(G)) = 1$; and since h maps $[0,1]$ *onto* $[0,2]$, we have $\overline{\lambda}(h(C)) = 1$.

Just as in Proposition 15.1 we can argue that there exists a subset S of $h(C)$ that does not belong to $\mathcal{L}$. The set $B := h^{-1}(S) \subseteq C$ is a subset of the $\bar{\lambda}-$null CANTOR set, so it is a $\bar{\lambda}-$negligible set; but the measure space $(\mathbb{R}, \mathcal{L}, \bar{\lambda})$ is complete, so $B \in \mathcal{L}$ and $\bar{\lambda}(B) = 0$. However, B cannot be a BOREL set, because otherwise we would have $S = h(B) \in \mathcal{B}(\mathbb{R}) \subseteq \mathcal{L}$, an impossibility, since S does not belong to $\mathcal{L}$. □

[51] The two results that follow can be considered as a "modern" versions of the ancient ZENO Paradox; see SKYRMS (2012).

Exercise 15.1. Show that it is impossible to construct a set function $\mu : \mathcal{P}(\mathbb{R}^d) \to [0, \infty)$, defined on *all* the subsets of the Euclidean space $\mathbb{R}^d$, which

(i) is countably additive,
(ii) assigns measure $\mu(K) = 1$ to the unit-cube K of this space, and
(iii) satisfies $\mu(E) = \mu(F)$ for sets E, F that are "congruent" (i.e., can be transformed into each other by means of translation, rotation and/or reflection).

(*Hint:* Just take $d = 1$, and consider the sets E as in the proof of Proposition 15.1 and $\{E_q\}$ as in (15.7); then try to arrive at a contradiction similar to that of (15.8). As BANACH & TARSKI (1924) show, this impossibility persists, even when one replaces the requirement (i) of "countable additivity" by that of "finite additivity".)

15.3 Additional Examples of Singular Distributions

Continuous but singular, CANTOR-like, distribution functions, arise very naturally when one considers infinite series of simple, BERNOULLI-type independent random variables. We study a few of them, following BILLINGSLEY (1986).

Let us consider, for instance, a sequence of independent random variables $X_1, X_2, \cdots$ with common distribution

$$\mathbb{P}(X_k = 1) = p_1 \in (0, 1)\,, \qquad \mathbb{P}(X_k = 0) = p_0 := 1 - p_1\,, \qquad \forall\ k \in \mathbb{N}\,.$$

What is the distribution function $F(x) = \mathbb{P}(X \leq x)$, $x \in \mathbb{R}$ *of the random variable* $X = \sum_{n\in\mathbb{N}} X_n\, 2^{-n}$ *?* We have encountered this situation already, in Remark 3.2.

Quite clearly, $F(0) = 0$ and $F(1) = 1$. For any given sequence $\mathfrak{d} = (d_1, d_2, \cdots) \in \{0, 1\}^{\mathbb{N}}$ we have

$$\mathbb{P}\big(X_k = d_k\,,\ k = 1, \cdots, n\big) \;=\; p_{d_1} \cdots p_{d_n} \;=\; p_1^{S_n(\mathfrak{d})}\, p_0^{n - S_n(\mathfrak{d})}\,,$$

where $S_n(\mathfrak{d}) := \sum_{k=1}^{n} d_k$. This expression tends to zero as $n \to \infty$, since we have either $\lim_{n\to\infty} S_n(\mathfrak{d}) = \infty$ or $\lim_{n\to\infty} (n - S_n(\mathfrak{d})) = \infty$, if not both. We deduce $\mathbb{P}\big(X_k = d_k\,,\ k \in \mathbb{N}\big) = 0$.

Any given $x \in [0, 1)$ can have at most two dyadic expansions of the type $x = \sum_{n\in\mathbb{N}} d_n\, 2^{-n}$, which implies $\mathbb{P}(X = x) = 0$ and the continuity of the distribution function $F(\cdot)$. As we did in Example 7.6, we shall work from now onward with the terminating expansion.

On the other hand, for every integer $0 \leq k < 2^n$ there is a unique terminating expansion $k\, 2^{-n} = \sum_{j=1}^{n} d_j\, 2^{-j}$, for some $(d_1, \cdots, d_n) \in \{0, 1\}^n$, and we get

$$\begin{aligned} F\big((k+1)\, 2^{-n}\big) - F\big(k\, 2^{-n}\big) \;&=\; \mathbb{P}\big(k\, 2^{-n} < X \leq (k+1)\, 2^{-n}\big) \\ &=\; \mathbb{P}\big(X_k = d_k\,,\ k = 1, \cdots, n\big) \;=\; p_{d_1} \cdots p_{d_n} \;>\; 0\,; \end{aligned} \tag{15.9}$$

in other words, $F(\cdot)$ is strictly increasing over the unit interval.

If $p_1 = p_0 = 1/2$ the expression on the right-hand side of (15.9) becomes 2^{-n}, and passing to the limit as $n \to \infty$ yields $F'(\cdot) \equiv 1$, $F(x) \equiv x$ on (0,1): the distribution of X in this case is uniform on the interval [0,1).

If $p_1 \neq 1/2$, consider for every $x \in [0,1)$ and $n \in \mathbb{N}$ an integer k_n so that $x \in K_n := \big(k_n\, 2^{-n},\ (k_n+1)\, 2^{-n}\big]$, and note

$$\frac{\mathbb{P}(X \in K_n)}{\lambda(K_n)} \;=\; 2^n \cdot F\big((k+1)\, 2^{-n}\big) - F\big(k\, 2^{-n}\big) \;\longrightarrow\; F'(x)\,, \qquad \text{as } n \to \infty\,.$$

We claim that $F'(x) = 0$, that is, the continuous, strictly increasing distribution function $F : [0,1] \to [0,1]$ is singular.

Let us argue this by contradiction: Suppose $F'(x) > 0$; then for two consecutive intervals K_n and K_{n+1}, the above relation gives

$$\frac{\mathbb{P}(X \in K_{n+1})}{\mathbb{P}(X \in K_n)} \;=\; \frac{1}{2} \cdot \frac{\mathbb{P}(X \in K_{n+1})}{\lambda(K_{n+1})} \cdot \frac{\lambda(K_n)}{\mathbb{P}(X \in K_n)} \;\longrightarrow\; \frac{1}{2}\,, \qquad \text{as } n \to \infty\,. \tag{15.10}$$

If K_n consists of the reals with terminating binary expansions beginning with $d_1, \cdots d_n$, then $\mathbb{P}(X \in K_n) = p_{d_1} \cdots p_{d_n}$; but then K_{n+1} consists of the reals with terminating binary expansions beginning with $d_1, \cdots d_n,\ d_{n+1}$ (where $d_{n+1} = 1$ if x lies to the right of the midpoint of the interval K_n, and $d_{n+1} = 0$ if to the left). Consequently,

$$\frac{\mathbb{P}(X \in K_{n+1})}{\mathbb{P}(X \in K_n)} \;=\; p_{d_{n+1}} \;\in\; \{p_0,\, p_1\}\,,$$

contradicting (15.10) since we have assumed $p_1 \neq 1/2$.

The continuous, strictly increasing but singular probability distribution function $F(\cdot)$ satisfies the recursions

$$F(x) \;=\; p_0 \cdot F(2x)\,, \quad \text{for} \quad 0 \le x \le \frac{1}{2}\,; \tag{15.11}$$

$$F(x) \;=\; p_0 + p_1 \cdot F(2x-1)\,, \quad \text{for} \quad \frac{1}{2} \le x \le 1\,. \tag{15.12}$$

These imply that each of the segments of the graph of $F(\cdot)$, below and above the 50% level, is identical to the whole graph, apart from changes in scale ("self-similar" or "fractal" behavior); see the picture on page 362 of BILLLINGSLEY (1979).

Exercise 15.2. Show the recursions of (15.11), (15.12).

Exercise 15.3. Let the random variables $X_1,\, X_2,\, \cdots$ be independent with common distribution $\mathbb{P}(X_k = i) = p_i\,,\ i = 0, 1, \cdots, r-1$ for all $k \in \mathbb{N}$. Here $p_0,\, p_1, \cdots, p_{r-1}$ are nonnegative numbers adding up to one.

Show that the distribution function $F(\cdot)$ of $X = \sum_{k \in \mathbb{N}} X_k\, r^{-k}$ is continuous; it is strictly increasing over the unit interval (0,1), if and only if all the p_i's are strictly positive. Show that $F(\cdot)$ is uniform over the unit interval (0,1) if $p_i \equiv 1/r$; and that $F(\cdot)$ is singular otherwise. What are the analogues of (15.11) and (15.12) in this case?

Exercise 15.4. Let $X_1,\, X_2,\, \cdots$ be independent random variables with common distribution

$$\mathbb{P}(X_k = 0) \;=\; \mathbb{P}(X_k = 2) \;=\; 1/2$$

for all $k \in \mathbb{N}$. Show that the series $W = \sum_{k\in\mathbb{N}} X_k \, 3^{-k}$ converges almost surely, and that the distribution of the resulting random variable is given by the CANTOR function $F(\cdot)$ of this section. (Recall, in this context, Exercise 7.9.)

Can you compute the moments, for instance the mean and the variance, of this distribution? its moment generating function?

16 Appendix: Elements of BANACH and HILBERT Spaces

We collect in this Appendix the basic definitions and properties of the abstract Banach and Hilbert space, for use throughout the course. Let us start with a real vector space $\mathfrak{X}$, and consider a mapping $\mathfrak{X} \ni x \mapsto ||x|| \in [0, \infty)$ with the following properties:

(i) $||x+y|| \leq ||x|| + ||y||$ for all x, y in $\mathfrak{X}$,

(ii) $||\alpha x|| = |\alpha| \cdot ||x||$, for all $x \in \mathfrak{X}$ and $\alpha \in \mathbb{R}$,

(iii) $||x|| = 0$ if and only if $x = 0$.

Such a mapping is called a *norm* on $\mathfrak{X}$. It is clear that $\rho(x,y) := ||x-y||$ defines a metric on $\mathfrak{X}$. We say that $\mathfrak{X}$ is a BANACH *space* if it is complete in the topology induced by this metric; in other words, if every CAUCHY sequence $\{x_n\}_{n\in\mathbb{N}} \subseteq \mathfrak{X}$ converges to some $x \in \mathfrak{X}$.

The primary examples of BANACH spaces are the $\mathbb{L}^p(\mu)$ spaces of (5.6) with the norms

$$||f||_p := \left(\int_\Omega |f|^p \, d\mu \right)^{1/p} < \infty$$

for $1 \leq p < \infty$, and

$$||f||_\infty := \inf \{a \geq 0 \,|\, \mu(\{\omega \in \Omega : |f(\omega)| > a) = 0\}$$

of (5.11) for $p = \infty$, respectively. The completeness of these spaces is the subject of Theorem 5.6.

An *operator* on the BANACH space $\mathfrak{X}$ is a mapping $T : \mathfrak{X} \to \mathbb{R}$. For every such mapping, we define the quantity

$$||T|| := \sup_{\substack{x\in\mathfrak{X} \\ x\neq 0}} \frac{|T(x)|}{||x||} \tag{16.1}$$

and say that T is *bounded* if $||T|| < \infty$. An operator T on $\mathfrak{X}$ is called *continuous*, if we have $\lim_{n\to\infty} T(x_n) = T(x)$ whenever the sequence $\{x_n\}_{n\in\mathbb{N}} \subseteq \mathfrak{X}$ converges to $x \in \mathfrak{X}$: that is, if $||x_n - x|| \to 0$ as $n \to \infty$. We say that T is a *linear operator* on $\mathfrak{X}$ if it satisfies

$$T(\alpha x + \beta y) = \alpha\, T(x) + \beta\, T(y), \quad \text{for every } x, y \in \mathfrak{X} \text{ and } \alpha, \beta \in \mathbb{R}. \tag{16.2}$$

For a linear operator T, we have

$$||T|| = \sup_{\substack{x\in\mathfrak{X} \\ ||x||=1}} |T(x)|. \tag{16.3}$$

The set of bounded, linear operators on the BANACH space $\mathfrak{X}$ is denoted by $\mathfrak{X}^*$ and is called the *dual* of $\mathfrak{X}$. *It is a real vector space with* $||\cdot||$ *of* (16.3) *as its norm.* Note that *every* $T \in \mathfrak{X}^*$ *is automatically continuous,* since for any sequence $\{x_n\}_{n\in\mathbb{N}} \subseteq \mathfrak{X}$ converging to $x \in \mathfrak{X}$ we have $|T(x_n) - T(x)| \leq ||T|| \cdot ||x_n - x|| \longrightarrow 0$ as $n \to \infty$, by linearity and boundedness ($||T|| < \infty$ in (16.1)).

The most important BANACH spaces, on which the most refined analysis is possible, are the abstract HILBERT spaces that we take up now. Let us denote by $\mathcal{H}$ a real vector space; an *inner product* is a function $(x,y) \mapsto \langle x,y\rangle : \mathcal{H}\times\mathcal{H} \to \mathbb{R}$, such that

(i) $\langle \alpha x + \beta y, z\rangle = \alpha\langle x,z\rangle + \beta\langle y,z\rangle$ for all x, y, z in $\mathcal{H}$ and α, β in $\mathbb{R}$,
(ii) $\langle x,y\rangle = \langle y,x\rangle$, for all x, y in $\mathcal{H}$,
(iii) $\langle x,x\rangle > 0$, $\forall\ x \in \mathcal{H}\setminus\{0\}$.

It follows from (i) that $\langle 0,z\rangle = 0$, $\forall\, z \in \mathcal{H}$. The pair $\big(\mathcal{H}, \langle\cdot,\cdot\rangle\big)$ is called a *pre*-HILBERT *space*, and for every $x \in \mathcal{H}$ we define

$$||x|| := \sqrt{\langle x,x\rangle}\,.$$

CAUCHY-SCHWARZ INEQUALITY: *For every x, y in $\mathcal{H}$ we have*

$$|\langle x,y\rangle| \le ||x||\cdot||y|| \tag{16.4}$$

with equality if and only if y and x are either collinear (*that is, $y = \gamma x$ for some $\gamma \in \mathbb{R}$*), *or else one of them is zero.*

Proof: Obvious if $y = 0$; if not, we observe that the quadratic function

$$\mathbb{R} \ni t \mapsto \langle x+ty, x+ty\rangle = ||x||^2 + 2t\langle x,y\rangle + t^2\,||y||^2 \tag{16.5}$$

is non-negative; thus we have $\Delta = 4\,(|\langle x,y\rangle|^2 - ||x||^2||y||^2) \le 0$ for its discriminant, and the result follows.

TRIANGLE INEQUALITY: *For every x, y in $\mathcal{H}$ we have*

$$||x+y|| \le ||x|| + ||y||\,. \tag{16.6}$$

Proof: Just read (16.5) with $t = 1$, and use (16.4) to obtain the new inequality $||x+y||^2 = ||x||^2 + 2\langle x,y\rangle + ||y||^2 \le ||x||^2 + 2||x||\,||y|| + ||y||^2 = (||x|| + ||y||)^2$.

It develops that the recipe $||x|| := \sqrt{\langle x,x\rangle}$, $x \in \mathcal{H}$ defines a *norm* on the pre-HILBERT space; if this latter is complete with respect to this norm, the resulting BANACH space is called a HILBERT *space*. We shall deal with such spaces from now on. The prime example comes from the space of square-integrable functions in integration theory.

Example 16.1. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a measure space and denote by $\mathbb{L}^2 \equiv \mathbb{L}^2(\mathbb{P})$ the space of measurable functions which are square-integrable: $\mathbb{E}(|X|^2) = \int |X|^2\, d\mathbb{P} < \infty$ (identifying functions that are equal $\mu-$a.e.). For any two elements f, g of this space, the product XY is an integrable function (since $2|XY| \le |X|^2 + |Y|^2$), and then

$$\langle X,Y\rangle := \mathbb{E}(XY) = \int XY\, d\mathbb{P}$$

defines an inner-product on the space. Completeness follows from Theorem 5.6, so $\mathbb{L}^2$ is a HILBERT space.

Exercise 16.1. If $x_n \to x,\ y_n \to y$ in $\mathcal{H}$, then $\langle x_n, y_n\rangle \to \langle x, y\rangle$ as $n \to \infty$.

We say that two elements x, y of $\mathcal{H}$ are *orthogonal,* if $\langle x, y\rangle = 0$. For any subset $A \subseteq \mathcal{H}$ of the HILBERT space, define the *orthogonal complement* of A by

$$A^\perp := \{x \in \mathcal{H} \,|\, \langle x, y\rangle = 0\,,\ \forall\, y \in A\}\,. \tag{16.7}$$

This is a *closed* subspace of $\mathcal{H}$, as follows easily from Exercise B.1; and $(A^\perp)^\perp$ is the smallest closed subspace of $\mathcal{H}$ that contains A.

Exercise 16.2. Show the *Parallelogram Identity*

$$||x+y||^2 + ||x-y||^2 \;=\; 2\,\left(||x||^2 + ||y||^2\right)\,,\quad \forall\ x,\, y \in \mathcal{H} \tag{16.8}$$

and the *Pythagorean Theorem*

$$\Big|\Big|\sum_{j=1}^{n} x_j\Big|\Big|^2 = \sum_{j=1}^{n} ||x_j||^2\,,\quad \text{when } \langle x_j, x_k\rangle = 0 \text{ for all } j \neq k\,. \tag{16.9}$$

Theorem 16.1. Projection in HILBERT Space: *If $\mathcal{G}$ is a closed subspace of the* HILBERT *space $\mathcal{H}$, then $\mathcal{H} = \mathcal{G} \oplus \mathcal{G}^\perp$. In other words, for every $h \in \mathcal{H}$ there exist $g_* \in \mathcal{G},\ z \in \mathcal{G}^\perp$ such that $h = g_* + z$.*

This decomposition is unique, and we have $||g_* - h|| = \inf_{g\in\mathcal{G}} ||g - h||$.

Proof: Fix $h \in \mathcal{H}$ and consider the distance $\delta := \inf_{g\in\mathcal{G}} ||g-h||$ of h from $\mathcal{G}$. We do not know yet whether the infimum is attained (this is part of what we shall try to prove), but consider a minimizing sequence $\{g_n\}_{n\in\mathbb{N}} \subseteq \mathcal{G}$ with $\lim_{n\to\infty} ||g_n - h|| = \delta$. From the parallelogram identity we get

$$\begin{aligned} ||g_n - g_m||^2 &= 2\left(||g_n - h||^2 + ||g_m - h||^2\right) - 4\,||(g_n+g_m)/2 - h||^2 \\ &\leq\; 2\left(||g_n - h||^2 + ||g_m - h||^2 - 2\delta^2\right) \longrightarrow 0 \end{aligned}$$

as $m, n \to \infty$. Thus the sequence $\{g_n\}_{n\in\mathbb{N}} \subseteq \mathcal{G}$ is CAUCHY, and converges (because $\mathcal{H}$ is complete) to some $g_* \in \mathcal{G}$ (because $\mathcal{G}$ is closed) with $||g_* - h|| = \delta$ (because we have $||g_* - h|| \geq \delta$ since $g_* \in \mathcal{G}$, and $||g_* - h|| \leq ||g_* - g_n|| + ||g_n - h|| \to \delta$ as $n \to \infty$).

We set now $z := h - g_*$, and claim $z \in \mathcal{G}^\perp$. To see this, fix an arbitrary $\gamma \in \mathcal{G}$ and observe

$$\delta^2 \;\leq\; ||h - (g_* - t\gamma)||^2 = ||z + t\gamma||^2 = \delta^2 + 2t\langle z, \gamma\rangle + t^2||\gamma||^2 =: f(t)\,,\quad t \in \mathbb{R}$$

(since $g_* - t\gamma \in \mathcal{G}$) and that the function $f(\cdot)$ attains its minimum at $t = 0$. Therefore $f'(0) = 0$, which implies $\langle z, \gamma\rangle = 0$, so $z \in \mathcal{G}^\perp$.

For any $z' \in \mathcal{G}^\perp$ with $z' \neq z$, the Pythagorean theorem gives $||h - z'||^2 = ||h - z||^2 + ||z - z'||^2 > ||h - z||^2$ (since $h - z \in \mathcal{G}$ and $z - z' \in \mathcal{G}^\perp$ are orthogonal); in particular, g_* is *the* (unique) element of $\mathcal{G}$ closest to h. If $h = g' + z' = g_* + z$ with $g' \in \mathcal{G},\ z' \in \mathcal{G}^\perp$, then $g' - g_* = z - z'$ belongs to $\mathcal{G} \cup \mathcal{G}^\perp = \{0\}$, and so $g_* = g',\ z = z'$. □

Now for any fixed $y \in \mathcal{H}$, the recipe $T_y(x) = \langle x, y\rangle,\ x \in \mathcal{H}$ defines a bounded linear operator $T_y : \mathcal{H} \to \mathbb{R}$ with norm

$$||T_y|| = \sup_{\substack{x\in\mathcal{H}\\ x\neq 0}} \frac{|\langle x, y\rangle|}{||x||} = ||y|| \tag{16.10}$$

in the notation of (16.1); indeed, the CAUCHY-SCHWARZ inequality gives $|T_y(x)| \le ||x|| \cdot ||y||$, with equality iff the vectors $x,\ y$ are either collinear ($y = \gamma x$ for some $\gamma \in \mathbb{R}$), or else $y = 0$. As a consequence, the mapping $\mathcal{T} : \mathcal{H} \to \mathcal{H}^*$ defined by $y \mapsto T_y(\cdot)\,,\ y \in \mathcal{H}$ introduces a *conjugate-linear isometry of* $\mathcal{H}$ *into its dual* $\mathcal{H}^*$ (the space of bounded, linear operators on $\mathcal{H}$). It turns out that this map is *bijective*, and that the HILBERT space $\mathcal{H}$ *is isometrically isomorphic* (via $\mathcal{T}$) *to its dual* $\mathcal{H}^*$ as the following result demonstrates.

Theorem 16.2. RIESZ Representation: *For any bounded linear operator* $T \in \mathcal{H}^*$, *there is a unique* $y \in \mathcal{H}$ *such that* $T(x) = T_y(x) \equiv \langle x, y\rangle\,,\ \forall\, x \in \mathcal{H}\,.$

Proof: If T is identically zero on $\mathcal{H}$, take $y = 0$. If not, observe that $\mathcal{G} = \{x \in \mathcal{H} \,|\, T(x) = 0\}$ is then a *proper, closed* subspace of $\mathcal{H}$.

Indeed, $\mathcal{G}$ is a vector space, because of the linearity of T; it is proper, because T is not identically equal to zero; and it is closed, because the convergence of the sequence $\{x_n\}_{n\in\mathbb{N}} \subseteq \mathcal{G}$ to $x \in \mathcal{H}$ and the continuity of the bounded, linear operator T implies $T(x) = \lim_{n\to\infty} T(x_n) = 0\,,$ thus $x \in \mathcal{G}$).

Now from Theorem 16.1 and the fact that $\mathcal{G}$ is a proper subspace of $\mathcal{H}$, there exists $z \in \mathcal{G}^\perp$ with $||z|| = 1$ (there exists $h \in \mathcal{H}$ with $T(h) \neq 0$, and $g_* \in \mathcal{G}\,,\ z \in \mathcal{G}^\perp \setminus \{0\}$ such that $h = g_* + z$; normalize, to get $||z|| = 1$). For any $x \in \mathcal{H}$ we have then $u := T(x)\, z - T(z)\, x \in \mathcal{G}$ (because $T(u) = T(x)T(z) - T(z)T(x) = 0$) as well as

$$0 = \langle u, z\rangle = T(x)\,||z||^2 - T(z)\,\langle x, z\rangle = T(x) - \langle x, y\rangle\,, \quad \forall\, x \in \mathcal{H}$$

with $y := T(z)\, z$. Uniqueness follows easily. □

Theorem 16.3. PARSEVAL, BESSEL, Totality: *Let* $\{\phi_n\}_{n\in\mathbb{N}}$ *be a sequence of orthonormal vectors in a Hilbert space* $\mathcal{H}$*:*

$$\langle \phi_n, \phi_k\rangle = 0 \quad \textit{and}\ ||\phi_n|| = 1\,, \quad \textit{for every}\ n \in \mathbb{N}\,,\ k \neq n\,.$$

Then the following statements are equivalent:

(i) *For any* $\psi \in \mathcal{H}$, *we have:* $\left|\left|\psi - \sum_{n=0}^{N} \langle \psi, \phi_n\rangle \phi_n\right|\right| \to 0$, *as* $N \to \infty$.

(ii) *For any* $\psi \in \mathcal{H}$, *we have the* PARSEVAL *Equation*

$$\sum_{n=0}^{\infty} \left|\langle \psi, \phi_n\rangle\right|^2 = ||\psi||^2\,. \tag{16.11}$$

(iii) *If* $\langle \psi, \phi_n\rangle = 0$ *for all* $n \in \mathbb{N}$, *then* $\psi = 0$ *(the Totality Property).*

(iv) *The sequence* $\{\phi_n\}_{n\in\mathbb{N}}$ *is a maximal sequence of orthonormal vectors in* $\mathcal{H}$.

(v) *The space of finite linear combinations of vectors from the sequence* $\{\phi_n\}_{n\in\mathbb{N}}$ *is dense in* $\mathcal{H}$.

Proof: We begin by observing that all orthonormal sequences $\{\phi_n\}_{n\in\mathbb{N}}$, all $\psi \in \mathcal{H}$, and all $N \in \mathbb{N}$ satisfy

$$0 \;\leq\; \Big|\Big| \,\psi - \sum_{n=0}^{N} \langle \psi, \phi_n \rangle \phi_n \,\Big|\Big|^2 \;=\; ||\psi||^2 - \sum_{n=0}^{N} \,|\langle \psi, \phi_n \rangle|^2 \,. \tag{16.12}$$

Thus (i) and (ii) are equivalent.

Next, (ii) clearly implies (iii). To see the reverse inplication, observe from (16.12) that, for all orthonormal systems $\{\phi_n\}_{n\in\mathbb{N}}$, we have the **BESSEL Inequality**

$$\sum_{n=0}^{\infty} \,\Big|\langle \psi, \phi_n \rangle\Big|^2 \;\leq\; ||\psi||^2 \,, \quad \forall\; \psi \in \mathcal{H} \,. \tag{16.13}$$

This implies that the left-hand side in (16.13) is a convergent series. Consequently, the sequence $\xi_N := \sum_{n=0}^{N} \langle \psi, \phi_n \rangle \phi_n \,,\; N \in \mathbb{N}$ converges in $\mathcal{H}$ to some vector $\xi := \sum_{n=0}^{\infty} \langle \psi, \phi_n \rangle \phi_n$, since

$$||\xi_N - \xi_M||^2 \;=\; \Big|\Big| \sum_{M+1}^{N} \langle \psi, \phi_n \rangle \phi_n \Big|\Big|^2 \;=\; \sum_{M+1}^{N} \,|\langle \psi, \phi_n \rangle|^2 \;\longrightarrow\; 0$$

as $M, N \to \infty$. The vector $\eta := \psi - \xi = \lim_{N\to\infty} (\psi - \xi_N)$ satisfies

$$\langle \eta, \phi_m \rangle \;=\; \lim_{N\to\infty} \left[\langle \psi, \phi_m \rangle - \sum_{n=0}^{N} \langle \psi, \phi_n \rangle \,\langle \phi_n, \phi_m \rangle \right] \;=\; 0 \,, \quad \forall\, m \in \mathbb{N} \,.$$

Thus, if (iii) holds, we have $\eta = 0$ or equivalently $||\psi - \xi_N||^2 \;\longrightarrow\; 0$ as $N \to \infty$, which then implies (16.11) by virtue of $(B.11)$. In other words, (iii) implies (ii) and we have established the equivalence of (i)-(iii). The equivalence of (iii) and (iv) is evident. Finally, (i) implies (v), while (v) implies (iii). The proof of the theorem is complete. □

An orthonormal sequence $\{\phi_n\}_{n\in\mathbb{N}}$ satisfying the conditions in Theorem 16 is called an *orthonormal basis* for $\mathcal{H}$. A HILBERT space admitting an orthonormal basis is called **separable.** We shall deal mostly with separable HILBERT spaces in this course.

Remark: It is possible to extend the notion of orthonormal basis to uncountable sets $\{\phi_\alpha\}_{\alpha\in I}$ of orthonormal vectors. The notion of convergence for an uncountable set can be easily extended, e.g., $\sum_{\alpha\in I} \psi_\alpha$ is said to *converge* to a given vector $\psi \in \mathcal{H}$, if, for all $\varepsilon > 0$, there exists a finite subset J of I such that $||\sum_{\alpha\in K} \psi_\alpha - \psi|| < \varepsilon$, for all subsets K of I, with $J \subseteq K$. The preceding discussion can be adapted in the same way. The axiom of choice guarantees then the existence of such a basis for *any* HILBERT space. However, this additional degree of generality is not necessary for the applications we shall consider in this course.

Remark: Under the conditions of Theorem we also have the "bilinear version" of the PARSEVAL equation (16.11):

$$\sum_{n=0}^{\infty} \,\langle \chi, \phi_n \rangle \,\langle \psi, \phi_n \rangle \;=\; \langle \chi, \psi \rangle \,, \quad \forall\; \chi \in \mathcal{H} \,,\; \psi \in \mathcal{H} \,. \tag{16.14}$$

PROOF OF THEOREMS 5.4 AND 5.5: (Adapted from DUDLEY (1989).) We shall provide a unified proof of these two results, in several steps.

• *Step 1: Reduction to the finite-measure case* $\mu(\Omega) + \nu(\Omega) < \infty$. By the assumptions of Theorem 5.4, we can write $\Omega = \cup_{n\in\mathbb{N}} E_n$, where $\{E\}_{n\in\mathbb{N}}$ are pairwiswe disjoint and $\mu(E_n) + \nu(E_n) < \infty$. For any measure ρ on $(\Omega, \mathcal{F})$, let $\rho_n(C) := \rho(C \cap E_n),\ n \in \mathbb{N}$ so that $\rho(C) = \sum_{n\in\mathbb{N}} \rho_n(C)$. Then $\rho(C) = 0$ is equivalent to $\rho_n(C) = 0,\ \forall\, n \in \mathbb{N}$; and we have $\nu < \mu \Leftrightarrow \nu_n < \mu,\ \forall\, n \in \mathbb{N} \Leftrightarrow \nu_n < \mu_n,\ \forall\, n \in \mathbb{N}$ as well as $\nu \perp \mu \Leftrightarrow \nu_n \perp \mu,\ \forall\, n \in \mathbb{N} \Leftrightarrow \nu_n \perp \mu_n,\ \forall\, n \in \mathbb{N}$.

On the other hand, for any sequence $\{\rho_n\}_{n\in\mathbb{N}}$ of measures on $(\Omega, \mathcal{F})$ with $\rho_n(E_n^c) = 0,\ \forall\, n \in \mathbb{N}$, the recipe $\rho(C) := \sum_{n\in\mathbb{N}} \rho_n(C)$ defines a measure for which $\rho < \mu \Leftrightarrow \rho_n < \mu_n,\ \forall\, n \in \mathbb{N}$ as well as $\rho \perp \mu \Leftrightarrow \rho_n \perp \mu_n,\ \forall\, n \in \mathbb{N}$. Thus, in proving Theorem 5.4, it is enough to assume that both measures $\mu,\ \nu$ are finite.

Now suppose we have proved Theorem 5.5 on each of the sets E_n, with some measurable $h_n : E_n \to [0, \infty),\ \forall\, n \in \mathbb{N}$. Then $h(\omega) \equiv h_n(\omega),\ \omega \in E_n$ defines a measurable function $h : \Omega \to [0, \infty)$ for which

$$\int_A h \,\mathrm{d}\mu = \sum_{n\in\mathbb{N}} \int_{A\cap E_n} h_n \,\mathrm{d}\mu_n = \sum_{n\in\mathbb{N}} \nu(A \cap E_n) = \nu(A) < \infty\,, \qquad \forall\, A \in \mathcal{F}\,,$$

by assumption of Theorem 5.5; in other words, $h \in \mathbb{L}^1(\mu) \cap \mathbb{L}^0_+$, and Theorem 5.5 then holds in all generality.

• *Step 2:* HILBERT*-space argument* (VON NEUMANN). Consider the HILBERT space $\mathcal{H} := \mathbb{L}^2 \equiv \mathbb{L}^2(\Omega, \mathcal{F}, \mu + \nu)$, and observe that the linear operator $\mathcal{H} \ni f \mapsto T(f) := \int_\Omega f\, d\nu \in \mathbb{R}$ is bounded, thus also continuous:

$$|T(f)| \le \int_\Omega |f| \,\mathrm{d}\nu \le \int_\Omega |f| \,\mathrm{d}(\mu + \nu) = ||f||_1 \le ||f||_2 \cdot \sqrt{\mu(\Omega) + \nu(\Omega)}\,,$$

from Exercise 5.7. Thus, from the RIESZ Representation Theorem $B.2$ (Appendix B), there exists a function $g \in \mathcal{H}$ such that for every $f \in \mathcal{H}$ we have

$$\int_\Omega f \,\mathrm{d}\nu = \int_\Omega fg \,\mathrm{d}(\mu + \nu)\,. \tag{16.15}$$

Remark: Note that, if indeed $\nu = \int h\, d\mu$ as postulated in the RADON-NIKODÝM theorem, then (16.15) becomes

$$\int_\Omega fh \,\mathrm{d}\mu \;=\; \int_\Omega fg \,\mathrm{d}\mu \;+\; \int_\Omega fgh \,\mathrm{d}\mu\,, \quad \forall\ f \in \mathcal{H}\,,$$

suggesting rather strongly that $h(1 - g) = g$ should hold $\mu-$a.e. and that we should choose $h = g/(1 - g)$ on $\{g \neq 1\}$. This observation will be very valuable below.

The equation (16.15) can be written equivalently, for every $f \in \mathcal{H}$, as

$$\int_\Omega f(1 - g) \,\mathrm{d}\nu = \int_\Omega fg \,\mathrm{d}\mu\,, \tag{16.16}$$

$$\int_\Omega f(1 - g) \,\mathrm{d}(\mu + \nu) = \int_\Omega f \,\mathrm{d}\mu\,. \tag{16.17}$$

It is easy to see that $0 \le g \le 1$, $(\mu+\nu)-$a.e. For if $(\mu+\nu)(\{g<0\})>0$, then (16.15) with $f=\mathbf{1}_{\{g<0\}}$ gives

$$\nu(\{g<0\}) = \int_{\{g<0\}} g\,\mathrm{d}(\mu+\nu) < 0\,;$$

and if $(\mu+\nu)(\{g>1\})>0$, then (16.17) with $f=\mathbf{1}_{\{g>1\}}$ gives

$$\mu(\{g>1\}) = \int_{\{g>1\}} (1-g)\,\mathrm{d}(\mu+\nu) < 0\,.$$

Both these conclusions are absurd.

In conclusion, we may assume that $0 \le g(\omega) \le 1$, $\forall\,\omega\in\Omega$; *then (16.15)-*(16.17) *hold for any non-negative, measurable* $f\in\mathbb{L}^0_+$, by the Monotone Convergence Theorem.

• *Step 3: Existence in Theorem 5.4*. Let us start by observing that the set $A=\{g=1\}$ has measure $\mu(A)=0$ (just read (16.16) with $f=\mathbf{1}_A$). In other words, we have $g<1$, $\mu-$a.e. Thus, for the two measures

$$\nu_s(E) := \nu(E\cap A)\,,\quad \nu_{ac}(E) := \nu(E\cap A^c)\,;\qquad E\in\mathcal{F} \tag{16.18}$$

we have $\nu=\nu_{ac}+\nu_s$, $\nu_s\perp\mu$ (because $\mu(A)=0$, $\nu_s(A^c)=\nu(\emptyset)=0$), as well as $\nu_{ac}<\mu$. This latter conclusion can be seen from the fact that, if $E\subseteq A^c=\{g<1\}$ has $\mu(E)=0$, then reading (16.16) with $f=\mathbf{1}_E$ gives $\int_E(1-g)\,d\nu = \int_E g\,d\mu = 0$, whence $\nu(E)=0$, $\nu_{ac}(E)=\nu(E)=0$.

• *Step 4: Uniqueness in Theorem 5.4*. Suppose that $\nu=\rho+\sigma$, where $\rho<\mu$, $\sigma\perp\mu$. Then $\rho(A\cap E)=0$ (because $\mu(A)=0$), whence

$$\nu_s(E) = \nu(E\cap A) = \rho(A\cap E)+\sigma(A\cap E) = \sigma(A\cap E) \le \sigma(E)$$

for any measurable set $E\in\mathcal{F}$; consequently $\nu_s\le\sigma$ and $\nu_{ac}-\rho=\sigma-\nu_s$ is a measure, both absolutely *and* singular with respect to μ. Thus, this measure is identically equal to zero: $\nu_{ac}\equiv\rho$, $\nu_s\equiv\sigma$.

• *Step 5: Proof of Theorem 5.5*. Now suppose $\nu\equiv\nu_{ac}$ $(\nu_s\equiv 0)$ and define a non-negative, measurable function as $h:=g/(1-g)$ on A^c, $h:=0$ on A. Then $f:=h\mathbf{1}_E$, $E\in\mathcal{F}$ is in $\mathbb{L}^0_+$ and we get $\int_E h(1-g)\,\mathrm{d}(\mu+\nu)=\int_E h\,\mathrm{d}\mu$, thus

$$\begin{aligned}
\int_E h\,\mathrm{d}\mu &= \int_{E\cap A^c} g\,\mathrm{d}(\mu+\nu) && \text{(from (16.17)}\\
&= \int_\Omega g\chi_{E\cap A^c}\,\mathrm{d}(\mu+\nu) \;=\; \nu(E\cap A^c) && \text{(from (16.15))}\\
&= \nu_{ac}(E) = \nu(E). && \text{(by assumption)}
\end{aligned}$$

The uniqueness follows easily. □

17 References

ALON, N. & SPENCER, J.H. (2000) *The Probabilistic Method*. Second Edition, J. Wiley & Sons, New York.

ANDERSON, G.W., GUIONNET, A. & ZEITOUNI, O. (2010) *An Introduction to Random Matrices*. Cambridge University Press.

APOSTOL, T.M. (1976) *Introduction to Analytic Number Theory*. Springer Verlag, New York.

ASH, R.B. (1970) *Basic Probability Theory*. J. Wiley & Sons, New York. Reissued in 2006 by Dover Publications, Mineola, NY.

BACHELIER, L. (1900) Théorie de la Speculation. *Annalles Scientifiques de l' École Normale Supérieure* **17**, 21-86.

BANACH, S. & TARSKI, A. (1924) Sur la décomposition des ensembles de points en parties respectivement congruents. *Fundamenta Mathematica* **6**, 244-277.

BELL, E.T. (1937) *Men of Mathematics: Volumes I and II.* Pelican Books, Melbourne.

BERNOULLI, J. (1713) *Ars Conjectandi*. Thurnisiorum, Basel.

BILLINGSLEY, P. (1968) *Convergence of Probability Measures*. J. Wiley & Sons, New York.

BILLINGSLEY, P. (1986) *Probability and Measure*. Second Edition, J. Wiley & Sons, New York.

BIRKHOFF, G.D. (1932) Proof of the ergodic theorem. *Proc. Nat'l. Acad. Sci. USA* **17**, 656-660.

BOLTHAUSEN, E. (1984) An estimate of the remainder in a combinatorial central limit theorem. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* **66**, 387-405.

BOREL, É. (1909) Les probabilités dénombrables et leurs applications arithmétiques. *Rendiconti Circolo Mat. Palermo* **27**, 247-271.

BRZEŹNIAK, Z. & ZASTAWNIAK, T. (1999) *Basic Stochastic Processess*. Springer Verlag, London.

CANTELLI, F.P. (1917) Su due applicazione di un teorema di G. Boole alla statistica matematica. *Rendiconti Accademia dei Linzei, Roma; Cl. Sci. Fis., Nat. (Ser. 5)* **26**, 39-45.

CARDANO, G. (1663) *Liber de Ludo Aleae*. [52] Translated as *"Book on Games of Chance"* by S.H. Gould, and published by Holt, Rinehart and Winston, New York (1953).

ČEBYŠEV, P.L. (1867) Des valeurs moyennes. *Journal des Mathématiques Pures et Appliquées* **12**, 177-184.

CHUNG, K.L. (1974) *A Course in Probability Theory*. Second Edition, Academic Press, New York.

COLDING, T. (2020) *Eilenberg Lectures.* Department of Mathematics, Columbia University, Spring 2020.

[52] Published posthumously. According to its author's autobiography *De Vita Propria*, this book was completed in 1525 when the author was still a young man, then re-written in 1565.

De FINETTI, B. (1937) La prévision: ses lois logiques, ses sources subjectives. *Ann. Inst. Henri Poincaré* **7**, 1-68.

De MOIVRE, A. (1730) *Miscellanea Analytica.* Paris.

DEVLIN, K. (2008) *The Unfinished Game: Pascal, Fermat, and the* 17^{th} *Century Letter That Made the World Modern*. Basic Books, New York.

DIACONIS, D. & SKYRMS, B. (2018) *Ten Great Ideas About Chance.* Princeton University press, Princeton, NJ.

DOOB, J.L. (1953) *Stochastic Processes*. J. Wiley & Sons, New York.

DUBINS, L.E. & SAVAGE, L.J. (1965) *How to Gamble if You Must: Inequalities for Stochastic Processes*. McGraw Hill Publishing Co., New York.

DURRETT, R. (2010) *Probability: Theory and Examples*. Fourth Edition, Cambridge University Press, New York.

DYSON, F. (1962) A Brownian-motion model for the eigenvalues of a random matrix. *J. Math. Phys.* **3**, 1191-1198.

DYSON, F. & MEHTA, M.L. (1963) Statistical theory for the energy levels of complex systems, IV. *J. Math. Phys.* **4**, 701-712.

ERDÖS, P. (1963) On a problem of graph theory. *Mathematical Gazette* **47**, 220-223.

EINSTEIN, A. (1905) Über die von der molekularkinetischen Theorie der Wärme geforderte Bewegung von in ruhenden Flüssigkeiten suspendierten Teilchen. *Annalen der Physik* **17**, 549-560.

ERDÖS, P. (1963) On a problem of graph theory. *Mathematical Gazette* **47**, 220-223.

ERDÖS, P. & KAC, M. (1940) The Gaussian law of errors in the theory of additive number-theoretic functions. *American Journal of Mathematics* **62**, 738-742.

ETEMADI, N. (1984) An elementary proof of the strong law of large numbers. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* **55**, 119-122.

FELLER, W. (1935) Über den zentralen Grenzwertsatz der Wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift* **40**, 521-559.

FELLER, W. (1968) *An Introduction to Probability Theory and Its Applications, Volume I*. Third Edition, J. Wiley & Sons, New York.

FELLER, W. (1971) *An Introduction to Probability Theory and Its Applications, Volume II*. Third Edition, J. Wiley & Sons, New York.

FISCHER, H. (2011) *An History of the Central Limit Theorem: From Classical to Modern Probability Theory*. Springer Verlag, New York.

FOLLAND, G.B. (1984) *Real Analysis: Modern Techniques and their Applications*. J. Wiley & Sons, New York.

GNEDENKO, B.V. (1976) *The Theory of Probability*. Mir Publishers, Moscow.

GRAHAM, L. & KANTOR, J.M. (2009) *Naming Infinity: A true story of religious mysticism and mathematical creativity.* Belknap Press. An imprint of Harvard University Press, Boston.

HUYGENS, Ch. (1657) *De Ratiociniis in Ludo Aleae*. Paris.

ITÔ, K. (1942) Differential Equations determining Markov Processes. (In Japanese.) *Zenkoku Shijō Sūgaku Danwakai* **1077**, 1352-1400.

ITÔ, K. (1944) Stochastic Integral. *Proc. Imperial Acad. Tokyo* **20**, 519-524.

ITÔ, K. & McKEAN, H.P. Jr. (1974) *Diffusion Processes and their Sample Paths*. Second Printing, Corrected. Springer Verlag, New York.

KAC, M. (1959) *Statistical Independence in Probability, Analysis and Number Theory*. Carus Mathematical Monographs **12**, Mathematical Association of America and J. Wiley & Sons, New York.

KARATZAS, I. & KARDARAS, C. (2021) *Portfolio Theory and Arbitrage.* To appear in *Graduate Studies in Mathematics,* American Mathematical Society, Providence, RI.

KARATZAS, I. & SHREVE, S.E. (1991) *Brownian Motion and Stochastic Calculus*. Second Printing, Corrected. Springer Verlag, New York.

KARDARAS, C. (2021) De Finetti's theorem and Bayesian updating. *Private Communication,* January 2021.

KARLIN, S. & TAYLOR, H.M. (1975) *A First Course in Stochastic Processes*. Second Edition, Academic Press, New York.

KOLMOGOROV, A.N. (1930) Sur la loi forte des grands nombres. *Comptes Rendus Acad. Sci. Paris* **191**, 910-912.

KOLMOGOROV, A.N. (1933) *Grundberiffe der Wahrscheinlichkeitsrechnung*. Springer Verlag, Berlin.

KOLMOGOROV, A.N. (1983) Combinatorial Foundations of Information Theory and the Calculus of Probabilities. *Russian Mathematical Surveys* **38**(4), 29-43.

KOLMOGOROV, A.N. & FOMIN, S.V. (1970) *Introductory Real Analysis.* Dover Publications, New York.

KOMOROWSKY, T. & LANDIM, C. & OLLA, S. (2012) *Fluctuations in Markov Processes.* Springer-Verlag, Berlin and Heidelberg.

LACEY, W. & PHILLIP, W. (1990) A note on the almost sure central limit theorem. *Statistics and Probability Letters* **1990**, 201-205.

LAPLACE, S.P. (1812) *Théorie Analytique des Probabilités*. Paris.

LÉVY, P. (1925) *Calcul des Probabilités*. Gauthier-Villars, Paris.

LÉVY, P. (1948) *Processus Stochastiques et Mouvement Brownien*. Gauthier-Villars, Paris.

LIGGETT, T.M. (1985) *Interacting Particle Systems*. Springer Verlag, New York.

LINDEBERG, J.W. (1922) Eine neue Herleitung des Exponentialgesetzes der Wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift* **15**, 211-225.

LINDVALL, T. (1977) A probabilistic proof of Blackwell's renewal theorem. *Annals of Probability* **5**, 482-485.

POISSON, S.D. (1837) *Recherches sur la Probabilité des Jugements.* Paris.

MANDELBROT, B. (1982) *The Fractal Geometry of Nature*. W.H. Freeman & Co., New York.

MARKOV, A.A. (1913) *Calculus of Probabilities*. St. Petersburg.

MARTIN-LÖF, P. (1966) The definition of random sequences. *Information and Control* **9**, 602-619.

NATANSON, I.P. (1955). *Theory of Functions of a Real Variable*. Frederick Ungar Publishing Company, New York.

NORRIS, J.R. (1997) *Markov Chains*. Cambridge University Press.

OLSHEN, R. (1974) A note on exchangeable sequences. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* **28**, 317-321.

PARTHASARATHY, K.R. (1967). *Probability Measures on Metric Spaces*. Academic Press, New York.

RESNICK, S.I. (1999) *A Probability Path*. Birkhäuser, Boston.

ROBBINS, H.E. (1955) A remark on the Stirling formula. *American Mathematical Monthly* **62**, 26-29.

ROBERTS, A.W. & VARBERG, D.E. (1973). *Convex Functions*. Academic Press, New York.

SHAFER, G. & VOVK, V. (2019) *Game-Theoretic Foundations for Probability and Finance.* J. Wiley & Sons, New York.

SHANNON, C.E. & WEAVER, W. (1949) *The Mathematical Theory of Communication*. University of Illinois Press, Urbana.

SHIRYAYEV, A.N. (1989) *Probability*. Second Edition, Springer-Verlag, New York.

SKYRMS, B. (2012) *From Zeno to Arbitrage: Essays on Quantity, Coherence and Induction.* Oxford University Press.

SOLOVAY, R.M. (1970) A model of set theory in which every set of reals is Lebesgue measurable. *Annals of Mathematics* **92**, 1-56.

SPITZER, F. (1969) Random processes defined through the interactions of an infinite particle system. *Lecture Notes in Mathematics* **89**, 201-223. Springer Verlag, New York.

STIRZAKER, D. (2003) *Elementary Probability*. Second Edition, Cambridge University Press.

TURÁN, P. (1934) On a theorem of Hardy and Ramanujan. *Journal of the London Mathematical Society* **9**, 274-276.

VILLE, J. (1939) *Étude Critique de la Notion du Collectif.* Gauthiers-Villars, Paris.

VON MISES, R. (1919) Grundlagen der Wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift* **5**, 52-99.

WALSH, J.B. (2012) *Knowing the Odds: An Introduction to Probability*. American Mathematical Society, Providence.

WIGDERSON, A. (2019) *Mathematics and Computation*. Princeton University Press, Princeton, NJ.

WIENER, N. (1923) Differential space. *J. Math. Physics* **2**, 131-174.

WIENER, N. (1924) Un problème de probabilités dénombrables. *Bull. Sc. Math. France* **52**, 569-578.

WIGNER, E.P. (1958) On the distribution of the roots of certain symmetric matrices. *Annals of Mathematics* **67**, 325-327.

WILLIAMS, D. (1991) *Probability with Martingales.* Cambridge Mathematical Texts. Cambridge University Press.

WISHART, J. (1928) The generalized product moment distribution in samples from a multivariate normal population. *Biometrica* **20A**, 32-52.

YOUNG, L.C. (1928) *The Theory of Integration*. Cambridge Tracts in Mathematics and Mathematical Physics, Vol. 21. Cambridge University Press, London.